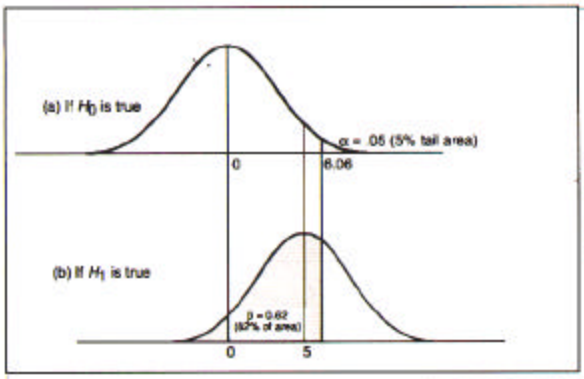
United States
Environmental Protection
Agency

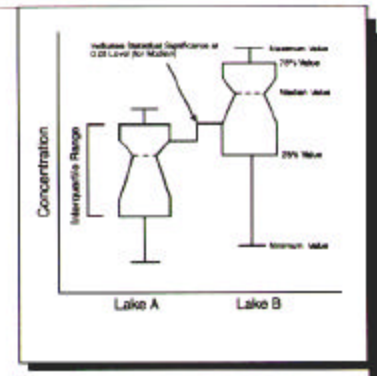
Office of Water
4304

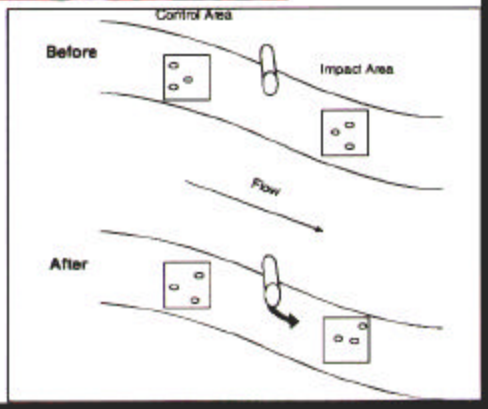
EPA 822-B-97-002
December 1997

EPA

Biological Criteria: Technical Guidance For Survey Design and Statistical Evaluation of Biosurvey Data







BIOLOGICAL CRITERIA
Technical Guidance for Survey Design and Statistical Evaluation of Biosurvey Data

Prepared for EPA by TetraTech, Inc.
Principal authors: Kenneth H. Reckhow, Ph.D. and William Warren-Hicks, Ph.D.

George Gibson, Jr., Ph.D.
Office of Science and Technology
Project Leader

Health and Ecological Criteria Division
Office of Water
U.S. Environmental Protection Agency
Washington, D.C. 20460

December 1997

Acknowledgements

This document was developed by the United States Environmental Protection Agency, Office of Science and Technology, Health and Ecological Criteria Division.

This text was written by Kenneth H. Reckow, PhD. And William Warren-Hicks, PhD. Jeroen Gerristen, PhD. of Tetra Tech, Inc. provided editorial and technical support. George R. Gibson, Jr., PhD. of USEPA was Project Leader and co-editor.

Disclaimer

This manual provides technical guidance to States, Indian Tribes, and other users of biological criteria to assist with survey design and statistical evaluation of biosurvey data. While this manual constitutes EPA's scientific recommendations regarding survey designs and statstical analyses, it does not substitute for the CWA or EPA's regulations; nor is it a regulation itself. Thus, it cannot impose legally binding requirements on the EPA, States, Indian Tribes, or the regulated community, and might not apply to a particular situation or circumstance. EPA may change this guidance in the future.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

FOREWORD

Biological Criteria: Technical Guidance for Survey Design and Statistical Evaluation of Biosurvey Data, by Kenneth H. Reckhow and William Warren-Hicks, was prepared for the U.S. Environmental Protection Agency to help states develop their biological criteria for surface waters and specifically to help water resource managers assess the reliability of their data. A good biological criteria program will be practical and cost effective, but above all it will be predicated on valid and scientifically sound information.

The application of the concepts and methods of statistics to the biological criteria process enables us ". . . to describe variability, to plan research so as to take variability into account, and to analyze data so as to extract the maximum information and also to quantify the reliability of that information" (Samuels, 1989).

This initial guidance document is intended to reintroduce statistics to the natural resources manager who may not be current in the application of this tool (and our ranks are legion, we just don't like to admit it). The emphasis is on the practical application of basic statistical concepts to the development of biological criteria for surface water resource protection, restoration, and management. Subsequent guides will be developed to expand on and refine the ideas presented here.

Address comments on this document and suggestions for future editions to George Gibson, U.S. Environmental Protection Agency, Office of Water, Office of Science and Technology (4304), 401 M Street, S.W., Washington, D.C. 20460.

CHAPTER 1 The Biological Criteria Program and Guidance Documents

Efforts to measure and manage water quality in the United States are an evolving process. Since its simple beginning more than 200 years ago, water monitoring has progressed from observations of the physical impacts of sediments and flotsam to chemical analyses of the multiple constituents of surface water to the relatively recent incorporation of biological observations in systematic evaluations of the resource. Further, although biological measurements of the aquatic system have been well-established procedures since the Saprobic system was documented at the turn of this century, such information has only recently been incorporated into the nation's approach to water resource evaluation, management, and protection.

The U.S. Environmental Protection Agency (EPA) is charged in the Clean Water Act (Pub. L. 100-4, §101) "to restore and maintain the chemical, physical, and biological integrity of the Nation's waters." To incorporate biological integrity into its monitoring program, the Agency established the Biological Criteria Program in the Office of Water.

This program provides technical guidance to the states for measuring biological integrity as an aspect of water resource quality. Biological integrity complements the physical and chemical factors already used to measure and protect the nation's surface water resources. Eventually all surface water types will be included in program technical guidance, including streams, rivers, lakes and reservoirs, wetlands, estuaries, and near coastal marine waters.

States will use this information to establish biological criteria or benchmarks of resource quality against which they may assess the status of their waters, the relative success of their management efforts, and the extent of their attainment or noncompliance with regulatory conditions or water use permits. These criteria are intended to augment, not replace, other physical and chemical methods, to help refine and enhance our water protection efforts.

The Concept of Biological Integrity

Biological integrity is the condition of the aquatic community inhabiting unimpaired waterbodies of a specified habitat as measured by community structure and function (U.S. Environ. Prot. Agency, 1990). Essentially, the concept refers to the naturally dynamic and diverse population of indigenous organisms that would have evolved in a particular area if it had not been affected by human activities. Such integrity or naturally occurring diversity becomes the primary reference condition or source of biological criteria used to measure and protect all waterbodies in a particular region.

Only the careful and systematic measuring of key attributes of the natural aquatic ecosystem and its constituent biological communities can determine the condition of biological integrity. These key attributes or biological endpoints indicate the quality of the waters of concern. They are established by biosurveys — by analyses based on the sampling of fish, invertebrates, plants, and other flora and fauna. Such biosurveys establish the endpoints or measures used to summarize several community characteristics such as taxa richness, numbers of individuals, sensitive or insensitive species, observed pathologies, and the presence or absence of essential habitat elements.

The careful selection and derivation of these measures (hereafter, metrics), together with detailed habitat characterization, is essential to translate the concept of biological integrity into useful biological criteria. That is, the quantitative distillation of the survey data makes it possible to compare and contrast several waterbodies in an objective, systematic, and defensible manner.

Narrative and Numeric Biological Criteria

Two forms of biological criteria are used in EPA's system of water resources evaluation and management.

■ **Narrative biological criteria** are general statements of attainable or attained conditions of biological integrity and water resource quality for a given use designation. They are qualitative statements of intent — promises formally adopted by the states to protect and restore the most natural forms of the system. Narrative criteria frequently include statements such as "the waters are to be free from pollutants of human origin in so far as achievable," or "to be restored and maintained in the most natural state." The statements must then be operationally defined and implemented by a designated state agency.

■ **Numeric criteria** are derived from and predicated on the same objective status as narrative criteria, which are then retained as preliminary statements of intent. The difference between the two is that the qualitative statement of integrity, the condition to be protected or restored, is refined by the inclusion of quantitative (numeric) endpoints as specific components of the criteria. Compliance with numeric criteria involves meeting stipulated thresholds or quantitative measures of biological integrity.

The formal adoption of criteria of either type into state law (with EPA concurrence) makes the criteria "standards." They are then applicable and enforceable under the provisions of the Clean Water Act.

Biological Criteria and Water Resource Management

Because these criteria will become the basis for resource management and possible regulatory actions, the manner of their design is of utmost importance to the states and EPA. The choice of metrics to represent and measure biological integrity is the responsibility of ecologists, biologists, and water resource managers. The Agency's role is to continue to develop technical guidance documents and manuals to assist in this process.

The purpose of this document is to present methods that will help managers interpret and gage the confidence with which the criteria can be used to make resource management decisions. Using this guidance, both the technician and the policymaker can objectively convert data into management information that will help protect water resources. However, the use and limits of the information must be clearly understood to ensure coordination and mutual cooperation between science and management.

An Overview of this Document

The focus of this document is on the basic statistical concepts that apply within the biocriteria program. From the program's inception, the problem statement, survey design, and the statistical methods used in the analysis must be correlated to provide functional results. Accordingly, chapter 2 begins with formulations of the problem statement — the focused objective that helps narrow the scope of observations in the ecosystem to those necessary to predict the status and impairment of the biota — and culminates in a discussion of hypothesis testing, the approach advocated in this guidance document. Chapter 2 also refers beginners to Appendix A for a succinct review of the basic statistics and statistical concepts used within the chapter and throughout this document.

Chapter 3 presents key issues associated with the design of the sample survey. Surveys are without doubt the critical element in an environmental assessment. Designs that minimize error, uncertainty, and variability in both biological and statistical measures have a great effect on decision makers. This chapter explores the difference between classical and experimental design and the issues involved with random, systematic, and stratified samples. Sample sizes and how to proceed in confusing circumstances round out the discussion.

Chapter 4 deals with problems that arise from hypothesis testing methods based on detecting the mean differences arising from two or more independent samples. The use and abuse of means testing procedures is an important topic. It should generally be keyed to the survey design, but other information should also be taken into consideration because errors of interpretation often involve assumptions about data.

Chapter 5 is a further discussion, with examples, of the basic concepts introduced in earlier chapters. Though hypothesis testing is generally preferred, this chapter discusses circumstances in which other procedures may be useful. It also introduces the role of cost-benefit assumptions in decision analysis and the limits of data collection and interpretation in the determination of causality. The reader should recall at all times the basic nature of this document. Advanced practitioners may look to the references used in preparing this document for additional options and discussion.

BIOLOGICAL CRITERIA
Technical Guidance for Survey Design and Statistical Evaluation of Biosurvey Data

CHAPTER 2 Classical Statistical Inference and Uncertainty

Before the biological survey can be designed and linked to statistical methods of interpretation, an exact formulation of the problem is needed to narrow the scope of the study and focus investigators on collecting the data. The choice of biological and chemical variables should be made early in the process, and the survey design built around that selection. Fancy statistics and survey designs may be appropriate, but biologically defined objectives should dominate and use the statistics, not the reverse (Green, 1979).

Formulating the Problem Statement

A clear statement of the objective or problem is the necessary basis on which the biological survey is designed. A general question such as "does the effluent from the municipal treatment plant damage the environment?" does little to help decision makers. Consider, however, their response to a more specific statement: "Is the mean abundance of young-of-the-year green sunfish caught in seines above the discharge point greater (with an error rate of 5 percent) than those similarly trapped downstream of the discharge point?" The precise nature of this question makes it a clear guide for the collection and interpretation of data.

The problem statement should minimally include the biological variables that indicate environmental damage, a reference to the comparisons used to determine the impact, and a reference to the level of precision (or uncertainty) that the investigator needs to be confident that an impact has been determined. In the preceding example, green sunfish are the biological indicator of impact, upstream and downstream seine data are the basis of comparison, and an error rate of 5 percent provides an acceptable level of uncertainty.

The problem statement, the survey design, and the statistical methods used to interpret the data are closely linked. Here, the survey design is an upstream/downstream set of samples with the upstream data providing a reference for comparison. A *t* test or rank sign test may be used to test for mean differences between the sites.

From a statistical standpoint, the biological variables (measures) used to show damage should have low natural variability and respond sharply to an impact relative to any sampling variability. Natural variability contributes to the uncertainty associated with their response to an impact. Lower natural variability permits reliable inferences with smaller sample sizes.

Examining historical data is an excellent means of selecting biological criteria that are sensitive to environmental impacts. Species that exhibit large natural spatial and temporal variations may be suitable indicators of environmental change only in small time scales or localized areas. If so, the use of such variables will limit the investigator's ability to assess environmental change in long-term monitoring programs. Historical data, combined with good scientific judgment, can be used to select biological criteria that exhibit minimal natural variability within the context of the site under evaluation.

Basic Statistics and Statistical Concepts

When a data set is quite small, the entire set can be reported. However, for larger data sets, the most effective learning takes place when investigators summarize the data in a few well-chosen statistics. The choice to trade some of the information available in the entire set for the convenience of a few descriptive statistics is usually a good one, provided that the descriptive statistics are carefully selected and correctly represent the original data.

Some descriptive statistics are so commonly used that we forget that they are but one option among many candidate statistics. For example, the mean and the standard deviation (or variance) are statistics used to estimate the center of a data set and the spread on those data. The scientist who uses these statistics has already decided that they are the best choices to describe the data. They work very well, for example, as representatives of symmetrically distributed data that follow an approximately normal distribution. Thus, their use in such circumstances is entirely justified. However, in other situations involving biological data, alternative descriptive statistics may be preferred.

Descriptive Statistics

Before selecting a descriptive statistic, the scientist must understand the purpose of the statistic. Descriptive statistics are often used in biological studies be-

cause the convenience of a few summary numbers outweighs the loss of information that results from not using the entire data set. Nevertheless, as much information as possible must be summarized in the descriptive statistics because the alternative may involve a misrepresentation of the original data.

The basic statistics and statistical techniques used in this chapter are further defined, described, and illustrated in the appendix to this document (Appendix A). Readers unfamiliar with descriptive statistics and graphic techniques should read Appendix A now and use it hereafter as a reference. Other readers may proceed directly to the tables in this chapter, which summarize the advantages and disadvantages of the statistical estimators and techniques described in the appendix.

The common measures of the center, or central tendency, of a data set are the mean, median, mode, geometric mean, and trimmed mean. None of these options is the best choice in all situations (see Table 2.1), yet each conveys useful information. The points raised in Table 2.1 are not comprehensive or absolute; they do, however, reflect the author's experience with these estimators.

Environmental contaminant concentration data are strictly positive, and sample data sets exhibit asymmetry (i.e., a few relatively high observations). Therefore, a transformation, in particular, the logarithmic transformation, should be applied to concentration and other data that exhibit these characteristics before analysis. When a transformation is used, data analysis and estimation occur within the transformed metric; if appropriate, the results may be converted back to the original metric for presentation.

A measure of dispersion — spread or variability — is another commonly reported descriptive statistic. Common estimators for dispersion are standard deviation, absolute deviation, interquartile range, and range. These estimators are defined, described, and illustrated with examples in the appendix; Table 2.2 summarizes when and how they may be used.

Table 2.3 summarizes four of the most useful univariate and bivariate graphic techniques, including histograms, stem and leaf displays, box and whisker plots, and bivariate plots. These methods are also illustrated in Appendix A.

Recommendations

There is no rigorous theoretical or empirical support for using the normal distribution as a population model for chemical and biological measures of water quality or as a model for errors. Instead, the evidence supports using the lognormal model. However, uncertainty about the correctness of the lognormal model suggests that prudent investigators will recommend estimators that perform well even if an assumed model is wrong.

Table 2.1—Measures of central tendency.

ESTIMATOR	ADVANTAGES	DISADVANTAGES	SHOULD CONSIDER FOR USE WHEN	SHOULD NOT USE WHEN
Mean	• Most widely known and used choice • Easy to explain	• Not resistant to outliers • Not as efficient[1] as some alternatives under deviations from normality	• Sample mean is required • Distribution is known to be normal • Distribution is symmetric	• Outliers may occur • Distribution is not symmetric
Median	• Easy to explain • Easy to determine • Resistant to others	• Not as efficient as the mean under normality	• Sample median is required • Outliers may occur	
Mode	• Easy to explain • Easy to determine	• Not as efficient as the mean under normality	• Most frequently observed value is required • Data are discrete or can be discretized	• More efficient options are appropriate
Geometric Mean	• Appropriate for certain skewed (lognormal) distribution	• Not as easy to explain as first three	• Distribution appears lognormal	• More widely known estimators are appropriate
Trimmed Mean	• Resistant to outliers	• Not as easy to explain as first three	• Outliers may occur and estimator efficiency is desired	

[1] Higher efficiency means lower standard error.

Table 2.2—Measures of dispersion.

ESTIMATOR	ADVANTAGES	DISADVANTAGES	SHOULD CONSIDER FOR USE WHEN	SHOULD NOT USE WHEN
Standard Deviation	• Most widely known • Routinely calculated by statistics packages	• Strongly influenced by outliers • Not as efficient[1] as some alternatives under even slight deviations from normality	• Sample standard deviation is required • Distribution is known to be normal	• Outliers may occur • Sample histogram is even slightly more dispersed than is a normal distribution
Median Absolute Deviation	• Resistant to outliers	• Not as efficient as the standard deviation under normality	• Outliers may occur	
Interquartile Range	• Resistant to outliers • Relatively easy to determine	• Not as efficient as the standard deviation under normality	• Outliers may occur	
Range	• Easy to determine	• Not as efficient	• Range is required	• Any of the above options is appropriate

[1] Higher efficiency means lower standard error.

Table 2.3—Useful graphic techniques.

TECHNIQUE	FEATURES	USEFUL FOR
Histogram	• Bar chart for data on a single (univariate) variable • Shows shape of empirical distribution	• Visual identification of distribution shape, symmetry, center, dispersion, and outliers
Stem and Leaf Display	• Same as histogram • Presents numeric values in display	• Same as histogram
Box and Whisker Plot	• Display of order statistics (extremes, quartiles, and median) • May be used to graph the same characteristic (e.g., variable) for several samples (e.g., different sampling sites)	• Visual identification of distribution shape, symmetry, center, dispersion, and outliers (single sample) • Comparison of several samples for symmetry, center, and dispersion
Bivariate Plot	• Scatter plot of data points (variable x versus variable y)	• Visual assessement of the strength of a linear relationship between two variables • Evidence of patterns, nonlinearity and bivariate outliers

Many books and articles have been written recently concerning the theoretical and empirical evidence in favor of nonparametric methods and robust and resistant estimators. Books that consider alternative estimators of center and dispersion (e.g., Huber, 1981; Hampel et al. 1986; Rey, 1983; Barnett and Lewis, 1984; Miller, 1986; Staudte and Sheather, 1990) build a strong case for more robust estimators than the mean and variance. Indeed, there is good evidence (Tukey, 1960; Andrews et al. 1972) that the mean and variance may be the worst choices among the common estimators for error-contaminated data. Several articles that involve comparisons of estimators on real data (e.g., Stigler, 1977; Rocke et al. 1982; Hill and Dixon, 1982) also favor robust estimators over conventional alternatives.

As a consequence, the median and the trimmed mean are recommended for the routine calculation of a data set's central tendency. The interquartile range and the median absolute deviation are recommended for calculation of the dispersion. These suggestions represent a compromise between robustness, ease of explanation, and calculation simplicity. For the trimmed mean, recommended amounts of trimming range from 10 percent (Stigler, 1977) to over 20 percent (e.g., Rocke et al. 1982). A critical argument in support of the trimmed mean is that interval estimation and hypothesis testing are still possible using the

t statistic (Tukey and McLaughlin, 1963; Dixon and Tukey, 1968; Gilbert, 1987).

Uncertainty

In statistics, uncertainty is a measure of confidence. That is, uncertainty provides a measure of precision — it assigns the value of scientific information in ecological studies. Scientific uncertainty is present in all studies concerning biological criteria, but uncertainty does not prevent management and decision making. Rather, uncertainty provides a basis for selecting among alternative actions and for deciding whether additional information is needed (and if so, what experimentation or observation should take place).

In ecological studies, scientific uncertainty results from inadequate scientific knowledge, natural variability, measurement error, and sampling error (e.g., the standard error of an estimator). In the actual analysis, uncertainty arises from erroneous specification of a model or from errors in statistics, parameters, initial conditions, inputs for the model, or expert judgment.

In some situations, uncertainty in an unknown quantity (e.g., a model parameter or a biological endpoint) may be estimated using a measure of variability. Likewise, in some situations, model error may be estimated using a measure of goodness-of-fit (predictions versus observations) of the model. In many situations, a judicious estimate of uncertainty is the only option; in these cases, careful estimation is an acceptable alternative and methods exist to elicit these judgments from experts (Morgan and Henrion, 1990).

In many studies, uncertainty is present in more than one component (e.g., parameters and models), so the investigator must estimate the combined effects of the uncertainties on the endpoint. This exercise, called error propagation, is usually undertaken with Monte Carlo simulation or first-order error analysis.

The outcome of an uncertainty analysis is a probability distribution that reflects uncertainty on the endpoint. However, uncertainty analysis may not always be the most useful expression of risk. Other expressions of uncertainty, such as prediction, confidence intervals, or odds ratios are easier to understand and interpret. If important error terms are ignored when a probability statement is made, the investigator must report this omission. Otherwise, the probability statement is not representative, and the uncertainties are underestimated.

Since uncertainty provides a measure of precision or value, it can be used by decision makers to guide management actions. For example, in some cases the uncertainty in a biological impact may be too large to justify management changes. As a consequence, managers may defer action until additional monitoring data can be gathered rather than require pollutant discharge controls. If the uncertainty is large and the estimated costs of additional pollutant controls quite high, it may be wise either to defer action or to look for smaller, relatively less expensive abatement strategies for an interim period while the monitoring program continues.

Though environmental planners at national, state, and local levels have rarely considered uncertainty in their planning efforts, their work has been generally successful over the past 20 years. It is, however, certainly possible that more effective management — that is, less costly, more beneficial management — might have occurred if uncertainty had been explicitly considered.

If overall uncertainty is ignored, the illusion prevails that scientific information is more precise than it actually is. As a consequence, we are surprised and disappointed when biological outcomes are substantially different from predictions. Moreover, if we don't calculate uncertainty, we have no rational basis for specifying the magnitude of our sampling program or the resources (money, time, personnel) that should be allocated to planning. Thus, decisions on planning and analysis are more likely based on convention and whim than on the logical objective of reducing scientific uncertainty.

Statistical analysis is largely concerned with uncertainty and variability. Therefore, uncertainty is an important concept in this guidance manual. The analyses presented here and in subsequent chapters are based on particular measures of uncertainty, for example, confidence intervals. These measures are "statistics"; they reflect data, and are not always considered in the broader context of uncertainty — that is, as establishing the uncertainty in a quantity of interest. We will, however, consider these statistics in the broader sense, with concern for the theoretical issues raised in this section. Particularly given the small samples that often occur with biocriteria assessments, investigators should ask the following questions:

- Do the data adequately represent uncertainty?
- Are all important sources of uncertainty represented in the data?
- Should expert scientific judgment be used to augment or correct measures of uncertainty?
- If components of uncertainty are ignored because they are not included in the data, are conclusions or decisions affected?

Statistical analysis is not a rote exercise devoid of judgment.

Statistical Inference

Statistical inference is gained by two primary approaches: (1) interval estimation, and (2) hypothesis testing. Interval estimation concerns the calculation of a confidence interval or prediction interval that bounds the range of likely values for a quantity of interest. The end product is typically the estimated quantity (e.g., a mean value) plus or minus the upper and lower interval. The same information is used in hypothesis testing; however, in hypothesis testing, the end product is a decision concerning the truth of a candidate hypothesis about the magnitude of the quantity of interest.

In a particular problem, the choice between using interval estimation or hypothesis testing generally depends on the question or issue at hand. For example, if a summary of scientific evidence is requested, confidence intervals are apt to be favored; however, if a choice or decision is to be made, hypothesis tests are likely to be preferred.

Interval Estimation

Statistical intervals, whether confidence or prediction, may be based on an assumed probability model describing the statistic of interest, or they may require no assumption of a particular underlying probability model.

Hahn and Meeker (1991) note that the proper choice of statistical interval depends on the problem or issue of concern. As a rule, if the interval is intended to bound a *population* parameter (e.g., the true mean), then the appropriate choice is the confidence interval. If, however, the interval is to bound a *future* member of the population (e.g., a forecasted value), then the appropriate choice is the prediction interval. Another statistical interval less frequently used in ecology is the tolerance interval, which bounds a specified *proportion* of observations.

In conventional (classical, or frequentist) statistical inference, the statistical interval has a particular interpretation that is often incorrectly stated in scientific studies. For example, if a 95 percent statistical interval for the mean is 7 ± 2, it is not correct to say that there is a 95 percent chance that the true mean lies between 5 and 9." Rather, it is correct to say that 95 percent of the time this interval is calculated, the true mean will lie within the computed interval. Although it sounds awkward and not directly relevant to the issue at hand, this interpretation is the correct meaning of a classical statistical interval. In truth, once it is calculated, the interval either does or does not contain the true value. In classical statistics, the inference from interval estimation refers to the procedure for interval calculation, not to the particular interval that is calculated.

Hypothesis Testing

Biosurveys are used for many purposes, one of which is to assess impact or effect. Resource managers may want to assess, for example, the influence of a pollutant discharge or land use change on a particular area. The effect of the impact can be determined based on the study of trends over time or by comparing upstream and downstream conditions. In some instances, the interest is in magnitude of effect, but concern often focuses simply on the presence or absence of an effect of a specific magnitude. In such cases, hypothesis testing is usually the statistical procedure of choice.

In conventional statistical analysis, hypothesis testing for a trend or effect is often based on a point null hypothesis. Typically, the point null hypothesis is that no trend or effect exists. The position is presented as a "straw man" (Wonnacott and Wonnacott, 1977) that the scientist expects to reject on the basis of evidence. To test this hypothesis, the investigator collects data to provide a sample estimate of the effect (e.g., change in biotic integrity at a single site over time). The data are used to provide a sample estimate of a test statistic, and a table for the test statistic is consulted to estimate how unusual the observed value of the test statistic is if the null hypothesis is true. If the observed value of the test statistic is unusual, the null hypothesis is rejected.

Table 2.4—Possible outcomes from hypothesis testing.

STATE OF THE WORLD	DECISION	
	ACCEPT H_0	REJECT H_0
H_0 is True	Correct decision. Probability = $1 - \alpha$; corresponds to the *confidence level.*	Type I error. Probability = α; also called the significance level.
H_0 is Fale (H_1 is True)	Type II error. Probability = β	Correct decision. Probablity = $1 - \beta$; also called power.

In a typical application of parametric hypothesis testing, a hypothesis, H_0, called the null hypothesis, is proposed and then evaluated using a standard statistical procedure like the t test. Competing with this null hypothesis for acceptance is the alternative hypothesis, H_1. Under this simple scheme, there are four possible outcomes of the testing procedure: the hy-

pothesis is either true or false, and the test results can be accepted or rejected for each hypothesis (see Table 2.4).

The point null hypothesis is a precise hypothesis that may be symbolically expressed:

$$H_0: \theta_1 - \theta_2 = 0$$
$$H_1: \theta_1 - \theta_2 \neq 0$$

where θ is a parameter of interest. An example of a point null hypothesis in words is, "no change occurs in mean IBI after the new wastewater treatment plant goes on line." Symbolically, it is expressed as

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_1: \mu_1 - \mu_2 \neq 0$$

where μ_1 is the "before" true mean and μ_2 is the "after" true mean. The test of the null hypothesis proceeds with the calculation of the sample means, $\bar{x}_1$ and $\bar{x}_2$. In most cases, the sample means will differ as a consequence of natural variability or measurement error or both, so a decision must be made concerning how large this difference must be before it is considered too large to result from variability or error. In classical statistics, this decision is often based on standard practice (e.g., a Type I error of 0.05 is acceptable), or on informal consideration of the consequences of an incorrect conclusion.

The result of a hypothesis test can be a conclusion or a decision concerning the rejected hypothesis. Alternatively, the result can be expressed as a "*p*-value," which quantifies the strength of the data evidence in favor of the null hypothesis. The *p*-value is defined as the probability that "the sample value would be as large as the value actually observed, if H_0 is true" (Wonnacott and Wonnacott, 1977). In effect, the *p*-value provides a measure of how likely a particular value is, assuming that the null hypothesis is true. Thus, the smaller the *p*-value, the less likely that the sample supports H_0. This is useful information; it suggests that *p*-values should always be reported to allow the reader to decide the strength of the evidence.

Common Assumptions

Virtually all statistical procedures and tests require the validity of one or more assumptions. These assumptions concern either the underlying population being sampled or the distribution for a test statistic. Since the failure of an assumption can have a substantial effect on a statistical test, the common assumptions of normality, equality of variances, and independence are discussed in this section. We must ask, for example, to what extent can an assumption be violated without serious consequences? Or how should assumption violations be addressed?

■ **Normality.** A common assumption of many parametric statistical tests is that samples are drawn from a normal distribution. Alternatively, it may be assumed that the statistic of interest (e.g., a mean) is described by a normal sampling distribution. In either case, the key distinction between parametric and nonparametric (or distribution-free) statistical tests is that a probability model (often normal) is assumed.

Empirical evidence (e.g., Box et al. 1978) indicates that the significance level but not the power is robust or not greatly affected by mild violations of the normality assumption for statistical tests concerned with the mean. This finding suggests that a test result indicating "statistical significance" is reliable, but a "nonsignificant" result may be the result of a lack of robustness to nonnormality. The normality of a sample can be checked using a normal probability plot, chi square test, Kolmogorov-Smirnov test, or by testing for skewness or kurtosis; however, many biological surveys are not designed to produce enough samples to make these tests definitive.

Normality of the sampling distribution for a test statistic is important because it provides a probability model for interval estimation and hypothesis tests. In some cases, transformation of the data may help the investigator achieve approximate normality (or symmetry) in a sample, if normality is required. Since nonnegative concentration data cannot be truly normal, and since empirical evidence suggests that environmental contaminant data may be described with a lognormal distribution, the logarithmic transformation is a good first choice. Therefore, in the absence of contrary evidence, we recommend that concentration data be log-transformed prior to analysis.

■ **Equality of Variance.** A second common assumption is that when two or more distributions are involved in a test, the variances will be constant across distributions. Many tests are also robust to mild violations of this assumption, particularly if the sample sizes are nearly identical. To test this assumption, a *t* test (usually a two-tailed one) can be performed; see Snedecor and Cochran (1967) for an example, and Miller (1986) for interpretive results. Conover (1980) provides an alternative, namely, nonparametric tests of equality of variances. Note that if two means are being compared based on samples with vastly different variances, the differences of interest may be more fundamental than the difference between the means.

■ **Independence.** The assumption of greatest general concern is independence. Most statistical tests (parametric and nonparametric) require a random sample, or a sample composed of independent observations. Dependency between or among observations

in a data set means that each observation contains some information already conveyed in other observations. Thus, there is less new independent information in a dependent data set than in an independent data set of the same sample size. Because statistical procedures are often not robust to violation of the independence assumption, adjustments are generally recommended to address anticipated problems.

Dependence in a sample can result from spatial or temporal patterns, that is, from persistence through time and space. In most types of analyses, the assumption of independence refers to independence in the disturbances (errors). For example, in a time series with temporal trend and seasonal pattern, dependence or autocorrelation in the raw data series may exist because of a deterministic feature of the data (e.g., the time trend or seasonal pattern).

This type of autocorrelation poses no difficulty; it is addressed by modeling the deterministic features of the data and subtracting the modeled component from the original series. Of particular concern in testing for trend is autocorrelation that remains after all deterministic features are removed (i.e., errors that are in the disturbances). When this situation arises, an adjustment to the trend test is necessary. Reckhow et al. (1993) provide guidance and software.

A similar situation can occur in the estimation of a regression slope or a central tendency statistic such as the mean or trimmed mean. In such cases, the independence assumption refers to the errors, as estimated by the residuals, around the regression line or the mean. If persistence or dependence is found in the residuals, then the independence assumption is violated and corrective action is needed. Options to address this problem include using an effective sample size (Reckhow and Chapra, 1983), or generalized, least squares for regression (see Kmenta [1986] or any standard econometrics regression text).

If the investigator finds positive autocorrelation in the disturbances (i.e., if each disturbance is positively correlated with nearby disturbances in the series), confidence interval estimates will be too narrow and may lead to rejection of the null hypothesis. Autocorrelation in the disturbances is the most common and potentially the most troublesome of the causes of assumption violations.

The degree of autocorrelation is a function of the frequency of sampling; that is, a data set based on an irregular sampling frequency cannot be characterized by a single, fixed value for autocorrelation. For biological time series, stream data obtained more frequently than monthly may be expected to be autocorrelated (after trends and seasonal cycles are removed). Stream survey data based on less frequent sampling are less likely to exhibit sample autocorrelation estimates of significance.

Parametric Methods — the t Test

Parametric approaches involve a model (e.g., regression slope) for any deterministic features and a probability model for the errors. In some cases, the deterministic model will be a linear, curvilinear, or step function, while the model for the errors is typically a normal probability distribution with independent, identically distributed errors. In other cases, the deterministic model may simply be a constant (as it is when interest focuses on an "upstream/downstream" comparison between two sites), though the probability model may in all cases be a normal probability distribution. The *t* test is a typical parametric test.

Using the *t* test

A Student's *t* statistic:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \qquad \textbf{(2.1a)}$$

has a Student's *t* distribution (n-1 degrees of freedom); here, "x" is the mean of a random sample from a normal distribution with true mean μ and constant variance, s is the sample standard deviation, and n is the sample size. In addition, for two samples:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\frac{s_1 + s_2}{2}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad \textbf{(2.1b)}$$

also has a Student's *t* distribution (n_1+n_2-2 degrees of freedom); here, x_1 and x_2 are the sample means; s_1 and s_2 are the sample standard deviations; and n_1 and n_2 are the sample sizes. This distribution is widely tabulated, and it is commonly used in hypothesis testing and confidence interval estimation for a sample mean (one-sample test; Equation 2.1a) or a comparison of sample means (two-sample test; Equation 2.1b).

When Student's *t* distribution is used in a hypothesis test (a *t* test), it is assumed that samples are drawn from a normal distribution, the variances are constant across distributions, and the observations are independent. Of these assumptions, Box et al. (1978) have shown that the *t* test has limited robustness to violations of the first two (normality and equality of variances); however, problems will occur if the observations are dependent. The scientist should probably be concerned about the first two assumptions only in situations in which the two data sets have substantially different variances and substantially different sample sizes (see Snedecor and

Cochran [1967] for F test calculations to compare variances).

An attractive variation of the t statistic for use in situations where outliers are of concern was proposed by Yuen and Dixon (1973; see also Miller, 1986; and Staudte and Sheather, 1990). They created an outlier-resistant, or robust, version of the t statistic (Equations 2.1a and 2.1b) using a trimmed mean and a Winsorized standard deviation. For example, if a t statistic is used to compare the means of two populations, the robust (trimmed t) version is

$$t_{tri} = \frac{\bar{x}_{tri1} - \bar{x}_{tri2}}{s_w \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad (2.2)$$

where $\bar{x}_{tri}$ = trimmed mean for sample i

s_w = Winsorized standard deviation

n_i = number of observations in sample i

A Winsorized statistic is similar to a trimmed statistic. For trimming, observations are ordered from lowest to highest, and the k-lowest and k-highest are removed from the sample for the calculation of the k-trimmed statistic (e.g., trimmed mean). For k-Winsorizing, observations are ordered from lowest to highest, and the k-lowest and k-highest are not removed, but are reassigned the values of the lowest observation and the highest observation remaining in the trimmed sample. The following example illustrates this.

A sample of 10 IBI values is obtained for analysis:
9, 31, 26, 25, 34, 38, 33, 31, 28, 37

And ordered from lowest to highest:
25, 26, 28, 29, 31, 31, 33, 34, 37, 38.

The 10 percent-trimmed sample is
26, 28, 29, 31, 31, 33, 34, 37

The 10 percent-Winsorized sample is
26, 26, 28, 29, 31, 31, 33, 34, 37, 37.

If we were to calculate the 10 percent-trimmed t statistic in Equation 2.2 for this IBI sample, we would use: (1) the trimmed sample (eight observations) to calculate a mean, and (2) the Winsorized sample (10 observations) to calculate a standard deviation. For the two-sample comparison of means, the trimmed t statistic has $(1\text{-}2k)(n_1+n_2)-2$ degrees of freedom or, in the above example, 7 degrees of freedom (df). The trimmed t statistic is an attractive option that should be considered whenever outliers are a concern.

The parametric approach is appropriate and advantageous if the deterministic model is a reasonable characterization of reality and if the model for errors holds. In such cases, parametric tests should be more powerful than nonparametric or distribution-free alternatives. Thus, the assumption that deterministic and probability models are correct is the basis on which the superior performance of parametric methods rests. If the assumptions concerning these models are incorrect, then the results of the parametric tests may be invalid and distribution-free procedures may be more appropriate.

Nonparametric Tests — the W test

Distribution-free methods, as the name suggests, do not require an assumption concerning the particular form of the underlying probability model for the data generation process. An assumption of independence is, however, usually made; therefore, autocorrelation can be as serious a problem in nonparametric methods as it is for parametric and robust methods. Distribution-free tests are often based on rank (or order); the sample observations are arranged from lowest to highest. The Wilcoxon-Mann-Whitney test or W test is a typical distribution-free test.

Using the *W* test

The W test is a two-sample hypothesis test, designed to test the hypothesis that two random samples are drawn from identical continuous distributions with the same center (alternative hypothesis: one distribution is offset from the other, but otherwise identical). This test is often presented as an option to the two-sample t test that should be considered if the assumption of normality is believed to be seriously in error. The W test has its own statistic, which is tabulated in most elementary statistics textbooks (i.e., those with a chapter on nonparametric methods). However, for moderate to large sample sizes (e.g., $n >$ 15), the statistic is approximately normal under the null hypothesis, so the standard normal table can be used.

The scientist should consider the W test for any situation in which the two-sample t test may be used. Comparative studies of these two tests indicate that while the t test is robust to violations of the normality assumption, the W test is relatively powerful while not requiring normality. Situations that appear severely nonnormal might favor the W test; otherwise the t test may be selected. Some statisticians (e.g., Blalock, 1972) recommend that both tests be conducted as a double check on the hypothesis.

Unfortunately, violation of the independence assumption appears to be as serious for the W test as for the t test. If these tests are to be meaningful, the scientist must confirm independence or make other adjustments as noted in Reckhow et al. (1993).

In essence, the *W* test is used to determine if the two distributions under study have the same central tendency, or if one distribution is offset from the other. To conduct the *W* test, the data points from the samples are combined, while maintaining the separate sample identity. This overall data set is ordered from low value to high value, and ranks are assigned according to this ordering.

To test the null hypothesis of no difference between the two distributions ($f[x]$ and $g[x]$)

$$H_0: f(x) = g(x)$$

the ranks, R_i, for the data points in one of the two samples are summed:

$$W = \sum R_i \qquad (2.3)$$

The ranks should be specified as follows (Wonnacott and Wonnacott, 1977): Start ordering (low to high, or high to low) from the end (high or low) at which the observations from the smaller sample tend to be greater in number, and sum the ranks to estimate *W* from this smaller sample. This estimate keeps *W* small as it is reported in most tables. For either one-sided or two-sided tests, if ties occur in the ranks, then all tied observations should be assigned the same average rank.

Statistical significance is a function of the degree to which, under the null hypothesis, the ranks occupied by either data set differ from the ranks expected as a result of random variation. For small samples, the *W* statistic calculated in Equation 2.3 can be compared to tabulated values to determine its significance (see Hollander and Wolfe, 1973). For moderate to large samples (where total *n* from both samples > 15), *W* is approximately normal (if the null hypothesis is true). Therefore, the *W* statistic may be evaluated using a standard normal table with mean ($E[W]$) and variance (Var[*W*]):

$$E(W) = n_A(n_B + n_A + 1)/2 \qquad (2.4)$$

$$\text{Var}(W) = n_B n_A(n_B + n_A + 1)/12 \qquad (2.5a)$$

If there are ties in the data, then the variance may be calculated as

$$Var(W) = \frac{n_A n_B}{12}\left[n_A + n_B + 1 - \frac{\sum_{j=1}^{g} t_j(t_j^2 - 1)}{(n_A + n_B)(n_A + n_B - 1)} \right] \qquad (2.5b)$$

where t_j is the size (number of data points with the same value) of tied group *j*. The effect of ties is negligible unless there are several large groups ($t_j \geq 3$) in the data set.

These statistics are used to create the standard normal deviate:

$$z = \frac{W - E(W)}{(Var(W))^{0.5}} \qquad (2.6)$$

where: n_A, n_B = the number of observations in samples *A* and *B* ($n_A < n_B$).

Example — an IBI case study

IBI data have been obtained from upstream and downstream sites surrounding a wastewater discharge. Assume independence.

Upstream	33	34	2.5	3.7	39	45	49	47	45	44
Downstream	26	30	18	32	36	36	43	42	41	41

(a) Test the null hypothesis that the true difference between the upstream and downstream IBI means is zero, versus the alternative hypothesis that the downstream IBI mean is *lower* than the upstream IBI mean.

$$H_0: \mu_U - \mu_D = 0$$

$$H_1: \mu_U - \mu_D > 0$$

First, some basic statistics for each sample:

	SAMPLE MEAN	SAMPLE STANDARD DEVIATION
Upstream	39.8	7.57
Downstream	34.5	8.09

For a comparison of two means based on equal sample sizes, the *t* statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\frac{(s_1 + s_2)}{2}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{39.8 - 34.5}{\frac{7.57 + 8.09}{2}\sqrt{\frac{1}{10} + \frac{1}{10}}} = \frac{5.3}{7.83\sqrt{0.2}} = \frac{5.3}{3.5} = 1.51$$

At the 0.05 significance level, the one-tailed *t* statistic for 18 degrees of freedom is 1.73. Since 1.51 < 1.73, we cannot reject the null hypothesis (at the 0.05 level).

(b) Test the null hypothesis (see part a) using the 10 percent trimmed *t* (10 percent trimmed from each end).

$$t_{tri} = \frac{\bar{x}_{tri1} - \bar{x}_{tri2}}{\frac{(s_{w1} + s_{w2})}{2}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{40.5 - 35.5}{\frac{5.83 + 6.39}{2}\sqrt{\frac{1}{10} + \frac{1}{10}}} = \frac{5.0}{6.12\sqrt{0.2}} = 1.83$$

At the 0.05 significance level, the one-tailed *t* statistic for 14 degrees of freedom is 1.76. Since 1.83 1.76, we reject the null hypothesis (at the 0.05 level).

(c) Test the null hypothesis (see part a) using the *W* test.

ORDER IBI VALUES																				
Upstream	49	47	45	45	44					39	37			34	33				25	
Downstream						43	42	41	41			36	36			32	30	26		18
ORDER																				
Upstream	1	2	3.5	3.5	5					10	11			14	15				19	
Downstream						6	7	8.5	8.5			12.5	12.5			16	17	18		20

Here the separate samples have been combined for the purpose of rank ordering. The W test statistic can then be calculated from the ranks:

$$W = \sum R_{Up} = 1+2+3.5+3+5+10+11+14+15+19 = 84$$

$$E(W) = n_A(n_B + n_A + 1)/2 = 10(10+10+1)/2 = 105$$

$$Var(W) = (n_B n_A / 12)[n_B + n_A + 1 - \{\sum t^3 - 1)\} / (n_B + n_A)(n_B + n_A - 1)]$$

$$= [(10)(10)/12][10 + 10 + 1 - \{(2)(3) + (2)(3) + (2)(3)\} / (10 + 10)(10 + 10 - 1)] = 174.61$$

$$z = \frac{(W - E(W))}{(Var(W))^{0.5}} = \frac{84 - 105}{174.61^{0.5}} = -1.59$$

At the 0.05 significance level, the one-tailed z statistic is 1.65. Since $1.59 < 1.65$, we cannot reject the null hypothesis (at the 0.05 level.

A glance at the IBI values and ranks in this example indicates a difference between the two samples (box plots and histograms would provide further supporting evidence). At issue is whether this difference in the sample is a chance occurrence or an indication of a true difference between the sites. If we adopt the conventional 0.05 level for hypothesis testing, then the conclusions from the three tests are ambiguous. Still, we can say the following about both the site comparisons and the methods:

(i) The downstream site is slightly impacted. Even though only one of the three test results yielded significance (at the 0.05 level), all three were close, suggesting a slight difference between the sites.

(ii) For each site, the lowest IBI value (25 for upstream, and 18 for downstream) is influential, particularly on the standard deviation. As a consequence, for the conventional t test, the denominator in the t statistic is inflated and rejection of the null hypothesis is less likely. Note that the lowest IBI value for the upstream site (IBI = 25) also affects the distribution-free W test. This IBI value holds a high rank (19) for the upstream sample, and substantially affects the test result. If that single IBI value had been 27 instead of 25, we would have rejected the null hypothesis at the 0.05 level.

(iii) The trimmed t is resistant to unusual observations or outliers, and thus provides the best single indicator of difference between the sites as conveyed by the bulk of the data from each site.

Conclusions

In hypothesis testing, the conclusion to not reject H_0 (in effect, to accept H_0) should not be evaluated strictly on the basis of a, the probability of *rejecting* H_0 when it is true (Type I error; see Table 2.4). Instead, we must be concerned with β, the probability of accepting H_0 when it is false (Type II error). Unfortunately, β does not have a single value, but is dependent on the true (but unknown) value of the difference between population means and on the sample size, n. For a particular testing procedure and sample size, we can determine and plot a relationship between the true difference between means and β. This plot is called the operating characteristic curve.

To understand the issues concerning significance and power (α and 1-β), consider the null hypothesis in the IBI case study:

H_0: The population mean IBI at the upstream site is the same as the population mean IBI at the downstream site.

In addition, because of the wastewater discharge, consider the general alternative hypothesis:

H_A: The population mean IBI at the upstream site is higher than the population mean IBI at the downstream site.

If we adopt $\alpha = 0.05$ (the probability of rejecting H_0 when it is true; Type I error) as our significance level, then Figure 2.1a displays the sampling distribution for the mean under H_0 with 18 degrees of freedom. The horizontal axis in Figure 2.1 is the "difference between the means"; thus, the sampling distribution is centered at zero in Figure 2.1a (consistent with zero difference between means under H_0). The 0.05-significant tail area (the "rejection region") begins at 6.06, which means that the sample difference must be greater than or equal to 6.06 for us to reject H_0. Since the difference between the means in our sample IBI was only 5.3, we are inclined to accept the null hypothesis, based on the conventional t test.

(Note: to find the beginning of the tail area multiply the t statistic times the standard error. In this example, the t statistic is 1.73 [one-sided, 0.05 level, 18 degrees of freedom], and the standard error is 3.5. Thus, the tail area begins at [1.73][3.5] = 6.06.)

Now suppose that the following alternative hypothesis, H_1, is actually true for the sample IBI case:

H_1: The population mean IBI at the upstream site is higher by 5.0 than the population mean IBI at the downstream site.

In addition suppose that while H_1 actually is true, we propose a hypothesis test for H_0 based on the acceptance region in Figure 2.1a (i.e., accept H_0 if the

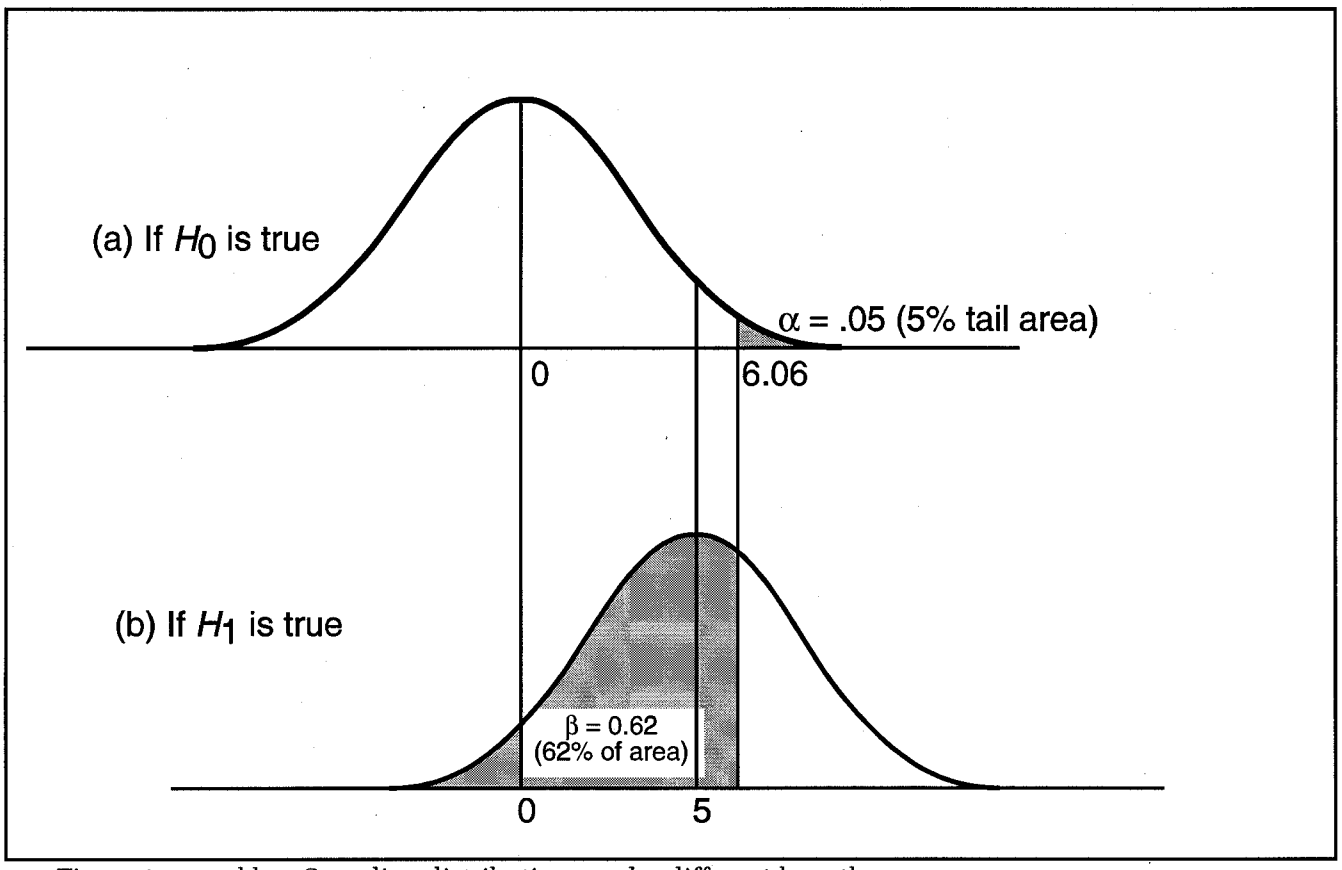


Figure 2.1a and b.—Sampling distributions under different hypotheses.

difference between the means is less than 6.06), which is exactly what occurred in our example. As we noted above, consideration of H_0 alone (Figure 2.1a) leads us to accept the null hypothesis.

Yet, with H_1 actually true (see Fig. 2.1b), if we propose a hypothesis test for H_0 based on the acceptance region in Figure 2.1a, there is a 62 percent chance that we will accept H_0 when it is actually false, according to Figure 2.1b (given the sample size in the example). This high likelihood of Type II error (see Table 2.4) underscores the danger of concluding the hypothesis test with acceptance of the null hypothesis. The power of this particular test is 1-β, or a 38 percent chance of detecting an IBI change of 5. Note that the specific alternative hypothesis H_1 is one example of an unlimited number of possibilities associated with the general alternative hypothesis H_A. Associated with H_1, β = 0.62 is one point on the power curve for this test and sample size. To properly determine the power of a test, we need to calculate β for a range of specific alternative hypotheses.

A second issue of concern in hypothesis testing is the problem of multiple simultaneous hypothesis testing, or "multiplicity" (Mosteller and Tukey, 1977). The classical interpretation of the 0.05 significance level (for α) associated with a hypothesis test is that 95 percent of the time this testing procedure is applied, the conclusion to accept the null hypothesis will not be in error if the null hypothesis is true. That is, on the average, one in 20 tests under these conditions will result in Type I errors.

The problem of multiplicity arises when an investigator conducts several tests of a similar nature on a set of data. If all but a few of the tests yield statistically insignificant results, the scientist should not ignore this in favor of those that are significant. The error of multiplicity results when one ignores the majority of the test results and cites only those that are apparently statistically significant. As Mosteller and Tukey (1977) note, the multiplicity error is technically the incorrect assignment of an α-level. When multiple tests of a similar nature are run on a set of data, a collective α should be used, associated with simultaneous test results. This tactic is typically referred to as the Bonferroni correction for correlation analysis.

The following comments from Wonnacott and Wonnacott (1972, pp. 201-202) summarize our attitude toward hypothesis testing:

> We conclude that although statistical theory provides a rationale for rejecting H_0, it pro-

vides no formal rationale for accepting H_0. The null hypothesis may sometimes be uninteresting, and one that we neither believe or wish to establish; it is selected because of its simplicity. In such cases, it is the alternative H_1 that we are trying to establish, and we prove H_1 by rejecting H_0. We can see now why statistics is sometimes called "the science of disproof." H_0 cannot be proved, and H_1 is proved by disproving (rejecting) H_0. It follows that if we wish to prove some proposition, we will often call it H_1 and set up the contrary hypothesis H_0 as the "straw man" we hope to destroy. And of course if H_0 is only such a straw man, then it becomes absurd to accept it in the face of a small sample result that really supports H_1.

Since there are great dangers in accepting H_0, the decision instead should often be simply to "not reject H_0," i.e., reserve judgment. This means that type II error in its worse form may be avoided; but it also means you may be leaving the scene of the evidence with nothing in hand. It is for this reason that either the construction of a confidence interval or the calculation of a prob-value is preferred, since either provides a summary of the information provided by the sample, useful to sharpen up your knowledge of what the underlying population is really like.

If, on the other hand, a simple accept-or-reject hypothesis test is desired, then we must look to a far more sophisticated technique. Specifically, we must explicitly take account not only of the sample data used in any standard hypothesis test (along with the adequacy of the sample size), but also:

1. *Prior belief.* How much confidence do we have in the engineering department that has assured us that the new process is better? Is their vote divided? Have they ever been wrong before?

2. *Loss involved in making a wrong decision.* If we make a type I error (i.e., decide to reject the old process in favor of the new, even though the old is as good), what will be the costs of retooling, etc.?

These comments amount to an advocacy of Bayesian decision theory. While it may be difficult to interpret a biosurvey in decision analysis terms, prior information and loss functions should, at a minimum, be considered in an informal manner. It is good engineering and planning practice to make use of all relevant information in inference and decision making.

BIOLOGICAL CRITERIA
Technical Guidance for Survey Design and Statistical Evaluation of Biosurvey Data

CHAPTER 3 Designing the Sample Survey

The design of the sample survey is a critical element in the environmental assessment process, and certain statistical methods are associated with specific designs. This chapter examines various types of survey design and shows how the selection of the design relates to the interpretation and use of data within the biocriteria program. For information on designs not covered in this chapter, see Cochran, 1963; Cochran and Cox, 1957; Green, 1979; Williams, 1978; and Reckhow and Stow, 1990.

Efforts to design sample surveys frequently result in situations that force the investigator to evaluate the trade-offs between an increase in uncertainty and the costs of reducing this uncertainty (Reckhow and Chapra, 1983). But major components of uncertainty, including variability, error, and bias in biological and statistical sources, can sometimes be controlled by a well-specified survey design.

For example, variability can be caused by natural fluctuations in biological indicators over space and time; error can be associated with inaccurate data acquisition or reduction; and bias can occur when the sample is not representative of the population under review or when the samples are not randomly collected. These sources of uncertainty should be evaluated before the sampling design is selected because the best design will minimize the effects of variability, error, and bias on decision making.

Critical Aspects of Survey Design

Data collection within the biocriteria program requires the investigator to address issues associated with both *classical* and *experimental* survey designs. In general, experimental survey design focuses on the collection of data that leads to the testing of a specific hypothesis. Classical survey design is motivated less by hypothesis testing than by the "survey" concept. That is, the investigator gathers a relatively small amount of data, the sample, and extrapolates from it a view of the totality of available information.

In this chapter, we will address issues that overlap these design types. In addition, we will focus on designs appropriate to local, site-specific situations. For larger geographic survey designs, see Hunsaker and Carpenter (1990), or Linthurst et al. (1986).

Variability

A critical aspect of sampling design is to identify and separate components of variability, including the important ones of time, space, and random errors. Yearly and seasonal variations and spatial variations like those caused by changes in geographic patterns should be accounted for in the survey design. A design that stratifies the sampling based on knowledge of spatial and temporal changes in the abundance and character of biological indicators is preferred to systematic random sampling. That is, if biological indicators are known to exhibit temporal and spatial patterns, then sampling locations and times must be adjusted to match the biological variability.

Representativeness and Sampling Techniques

The object of a biological survey design is to reduce the total information available to a small sample: observations are made and data collected on a relatively small number of biological variables. Representativeness is, therefore, a key consideration in the design of sample collection procedures. The data generated during the survey should be representative of the population or process under evaluation. Biased samples occur when the data are not representative of the population. For example, a sample mean may be low (biased) because the investigator failed to sample geographic areas of high abundance.

Several techniques can increase the odds of collecting a representative sample; however, the technique most frequently used is *random sampling*. Theoretically, in simple random sampling, every unit in the population has the same chance of being included in the sample. Random sampling is a physical way to introduce independence among environmental measurements. In addition, random sampling has the affect of minimizing various types of bias in the interpretation of results.

If the geographic area sampled is large, with known or suspected environmental patterns, a good technique is to divide the area into relatively homogenous sections and randomly sample within each one. This technique is known as *stratified sampling*. Samples can be allocated to each section in proportion to the size of the area or to the known abundance of organisms within each area. In still other cases, *systematic sampling* may be appropriate. Systematic sampling improves precision in the sample estimates, especially when known spatial patterns exist

(Cochran, 1963). Randomly allocated replicate samples collected on a grid allow for good spatial coverage of patchy environments, yet minimize the potential for sampling bias.

Cause and Effect

In classical statistical experiments, a population is identified and randomly divided into two groups. The treatment is administered to one group; the other group serves as the control. The difference in the average response between the two groups indicates the effect of the treatment, and the random assignment of individuals to the groups permits an inference of causality because the observed difference results from the treatment and not from some preexisting difference between the groups.

In an ecological assessment, the treatment and control groups are not selected at random from a larger population, since the impacted site cannot be selected at random. And no matter how carefully the reference site is matched, the investigator cannot compensate for the lack of random selection. In this sense, a statistically valid test of the hypothesis that an observed difference between an impacted site and a control site results from a specific cause is impossible. The hypothesis that the two sites are different can be tested, but the difference cannot be attributed to a specific cause. In statistical terms, the stress on the impacted site is completely confounded with preexisting differences between the impact and reference site.

Although a firm case can be made that a site is subject to adverse impacts, investigators must realize that the site is an experimental unit that cannot be replicated. They must take care to avoid "pseudoreplication" (Hurlbert, 1984) — the testing of a hypothesis about adverse effects without appropriate statistical design or analysis methods. The problem is a misunderstanding or misspecification of the hypothesis being tested. It is avoided by understanding that only the hypothesis of a difference between sites can be statistically tested. Cause-and-effect issues cannot be resolved using statistical methods. Of course, establishing that a difference exists is an essential step in the process of demonstrating an adverse ecological effect. If there is no detectable difference, there is no cause to establish.

Methods used to establish causality can make use of statistical techniques, such as regression or correlation. For example, regression can be used to show that toxicity increases along with the concentration of some chemical known to originate from a wastewater outfall. The regression describes the relationship; it does not imply the cause, though presence of a strong relationship is evidence that a link exists.

One way to resolve these issues is to collect both spatial and temporal data from a control site. If the spatial control is missing and only before and after impact samples are available at the impacted site, statistical tests cannot rule out the possibility that the change would have occurred with or without the impact. If the temporal control is missing, the statistical tests cannot rule out the possibility that the differences between the control and impact site may have occurred prior to the impact. In practice, control data are rarely available in both spatial and temporal dimensions. Therefore, most environmental assessments detect only that differences exist between the control and impact sites. The causal link is more difficult to discern.

Controls

In environmental assessments, control or reference data are used in hypothesis tests to evaluate whether data from the control and impact site are statistically different. Evidence of impact is based on changes in the biological community that did not occur in the control area. Sources of control information include baseline data, reference site data, and numeric standards. The case for causality can be strengthened if the controls are properly selected.

In an ideal study design, both temporal and spatial control data should be collected (Green, 1979). The control site should be geographically separated from the impacted site but have similar physical and ecological features (e.g., elevation, temperature, wind patterns, and habitat type and disturbance). In aquatic habitats, parameters such as stream order, flow rate, and stream hydrography should be considered. Ideally, biological indicators of impact should be collected at the control site before and after the impact occurs.

Statistically, a valid control site should have conservative properties. That is, its statistics should be the same as at the impacted site except for the effects of the impact. Physical, chemical, and ecological variables should be measured and statistically evaluated to confirm that the impact and control sites are properly matched. Investigators should test for mean differences as well as differences in distribution. In addition, the variance of the physical and ecological similarities between the control and impact sites should be the same over time. For example, if the mean pH between the two sites is consistent but the impact site experiences much wider swings in pH than the control site, then the ability to confidently detect an impact for a pH-dependent toxicant is compromised. Samples within the control and reference site should be randomly allocated at some level. For example, in a random sampling design (Fig. 3.1), the

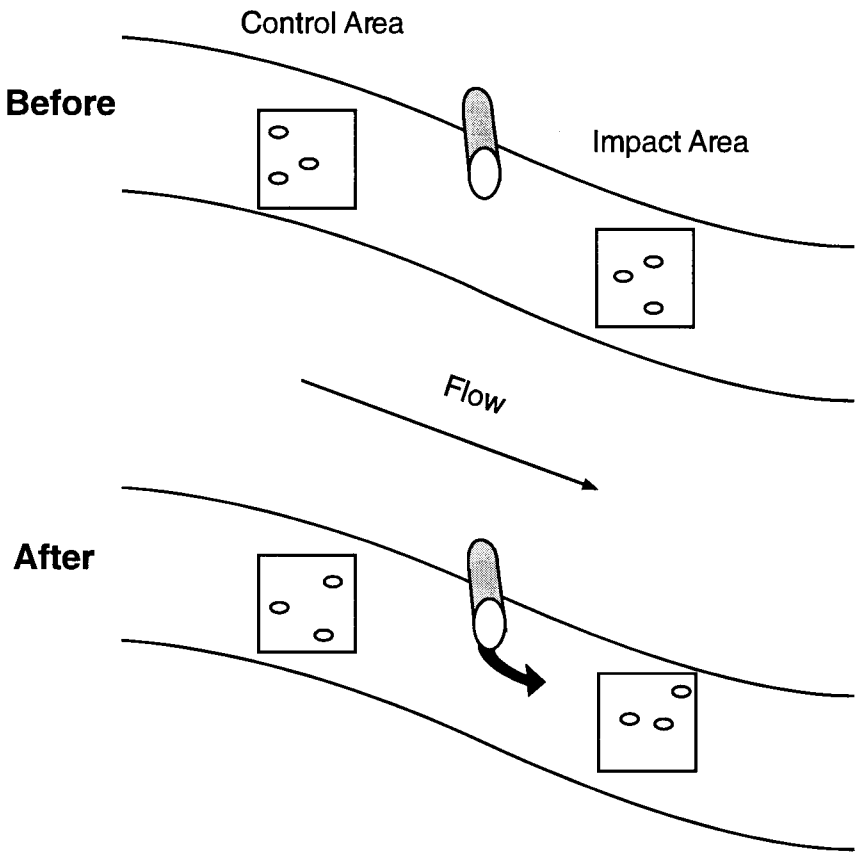


Figure 3.1—Random before and after control impact (BACI) sample design having both temporal and spatial dimensions. Random samples indicated are from within areas identified as being of similar habitat. (Adapted from Green, 1979.)

samples would be randomly allocated in a temporal/spatial framework that would allow for a number of different statistical analyses, including analysis of variance (ANOVA).

In an optimal study design, the impact would be in the future. Thus, baseline data providing a temporal control would be available to the investigator. In practice, baseline data are rarely available, and the investigator cannot be certain whether differences between the impact and control sites preceded or followed the impact. Therefore, cause and effect cannot be determined. However, the fact that a difference exists allows the investigator to hypothesize a causal link.

In some cases, biological variables collected at an impact site may be compared to a fixed numeric value rather than to a set of identical measurements collected at a reference site. Nevertheless, the issues associated with demonstrating causality remain the same. In addition, the investigator should note that the numeric criterion has no variance. It is usually presented as a single number with no associated uncertainty. In such cases, a *t* statistic (see chapter 4) would be appropriate. As an alternative to the numeric criterion, investigators could use the data from which the criterion was derived. Uncertainty estimates from that data set could be used in statistical comparisons.

Key Elements

Several specific survey designs are appropriate for use in a biocriteria program, but designs for a particular environmental assessment should be developed with the aid of a consulting statistician. Such plans should include the following key elements, beginning with the notion of a pilot study.

Pilot Studies

In a pilot study, the investigator makes a limited survey of the variables that determine impact at both the impact and control site. Data from the survey can be used to estimate sample sizes, evaluate sampling methods, establish important variance components, and critique or reevaluate the larger design. The sample size helps determine the particular levels of statistical confidence that can be gleaned from the study. In general, a pilot study can save time and effort by verifying an investigator's preliminary assumptions and initial evaluations of the impact site. Current studies

and historical data collected at the site of interest or similar sites can be used to help establish a good monitoring design.

Location of Sampling Points

A second key issue in the study design is the location of the sampling points. Many specific designs and variations are available, including (1) completely random sample designs, (2) systematic sample designs, and (3) stratified random sample designs.

■ **Random Samples**. In complete random sampling, every potential sampling point has the same probability of selection. The investigator randomly assigns the sample points within the impact site and independently within the control site. No attempt is made to partition the impact and control sites either spatially or temporally except to ensure similar physical habitats. The sampling units are numbered sequentially, and the selection is made using a random number table or computer-generated random numbers.

The advantage of random sampling is that statistical analysis of data from points located completely at random is comparatively straightforward. In addition, the method provides built-in estimates of precision. On the other hand, random sampling can miss important characteristics of the site, spatial coverage tends to be nonuniform, and some points may be of little interest.

■ **Systematic Samples**. Systematic sampling occurs when the investigator locates samples in a nonrandom but consistent manner. For example, samples can be located at the nodes of a grid, at regular intervals along a transect, or at equally spaced intervals along a streambank. The grid or interval can be generated randomly, after which the position of all samples is fixed in space.

Systematic sampling has two advantages over simple random sampling. First, it is easier to draw, since only one random number is required. Second, the sampling points are evenly distributed over the entire area. For this reason, systematic sampling often gives more accurate results than random sampling, particularly for patchy environments or environments with distinct discontinuous populations.

Systematic sampling also has its disadvantages. For example, if the magnitude of the biological variable exhibits a fixed pattern or cycle over space or time, then systematic sampling is unlikely to represent variance of the entire population. Suppose an organism has several hatches, roughly at equally spaced time intervals during the sampling period, then samples taken at fixed-time intervals may provide a biased estimate of the average number of individuals alive at one time. If possible, the population should be checked for such periodicity. If periodicity is found or suspected but not verifiable, systematic sampling should not be used.

Another disadvantage of systematic sampling is that it is more complicated to estimate the standard error than if random sampling had been used. Despite these problems, systematic sampling is often part of a more complex sampling plan in which it is possible to obtain unbiased estimates of the sampling errors.

■ **Stratified Random Samples**. Stratified samples combine the advantages of random and systematic sampling. Stratified random sampling consists of the following three steps: (1) the population is divided into a number of parts, called strata; (2) a random sample is drawn independently in each stratum, and (3) an estimate of the population mean is calculated. Thus:

$$\bar{y}_{st} = \frac{\sum N_h \bar{Y}_h}{N} \qquad (3.1)$$

where $\bar{y}_{st}$ is the estimate of the population mean, N_h is the total number of sampling units in the h^{th} stratum, and $\bar{y}_h$ is the sample mean in the h^{th} stratum, and $N = \sum N_h$ is the size of the population. Note that N_h are not sample sizes but the total sizes of the strata, which must be known to calculate this value.

Stratification is employed if it can be shown that differences between the strata means in the population do not contribute to the sampling error in the estimate of $\bar{y}_h$. In other words, the sampling error of $\bar{y}_h$ arises solely from variations among sampling units that are in the same stratum. If the strata can be formed so that they are internally homogeneous, a gain in precision over simple random sampling can occur.

In stratified sampling, the sample size can vary independently across strata. Therefore, money and human resources can be allocated efficiently across strata. As a general rule, strata with the greatest uncertainty (i.e., with the largest expected variance, or about which little is known) should receive the greatest amount of sampling effort.

For environments that are known to be fairly homogeneous with respect to the biological variable under consideration, stratified random sampling will not add precision to the population estimates. In fact, using stratification in these environments may introduce a loss of precision and a possible bias in the population estimates. In these cases, the investigator may save a great deal of time and effort by using simple random sampling in the sampling plan.

Location of Control Sites

Under EPA's biocriteria program, states may establish either site-specific reference sites or ecologically similar regional reference sites for comparison with impacted sites (U.S. Environ. Prot. Agency, 1990). Typical site-specific reference sites may be established along a gradient. For example, a reference site can be established upstream of a wastewater outfall (Fig. 3.1). Gradients work well for rivers and streams; for larger waterbodies, reference sites can be established on a one-to-one basis with a similar waterbody in the region not experiencing the impact under evaluation.

An important consideration in site-specific reference conditions is to establish that the control site is not impaired at all or that it is only minimally impaired. In particular, baseline data should be obtained to demonstrate that the impact is linked to the differences detected between the reference site and the control site.

Ideally, a reference site should exhibit no impairment; however, natural variability in biological data may make the determination of minimal or no impact difficult, especially if the impact is relatively small. An interesting method for site selection is to establish several reference sites based on their physical similarities with the impact site. For example, selecting one reference site with higher flow than the impact site and another with lower flow may increase the investigator's ability to determine the presence of a real impact. Comparisons of data collected from the impact and reference sites should provide consistent interpretations of the impact, regardless of which reference site is used in the comparison.

Minimizing temporal variation in biological measurements can be critical to the evaluation of control and impacted sites. A general rule is that samples should be obtained from the control and reference sites during the same time periods. It may be feasible to target an index period (e.g., late spring or summer) in which the biological variables are assumed to be appropriate indicators of ecological health (e.g., the period of maximum abundance or the period of minimum variation in water chemistry). However, for some organisms, periods of maximum abundance may also be periods of high variability. In this case, periods of low abundance but stable conditions can be used to help the investigator detect impairment if it exists.

Estimation of Sample Size

A final key component in developing a survey design is to determine how many samples are required. In most plans, the issue involves a trade-off between the accuracy of the sample estimate and the magnitude of available monetary and human resources. Consequently, the first step is to determine how large an error can be tolerated in the sample estimate. This decision requires careful thought; it depends on how the collected data will be used and the consequences of a sizable uncertainty associated with the sample estimates. Thus, in reality, selecting a sample size is somewhat arbitrary and driven by practical considerations of time and money. Investigators should, however, always approach the selection of sample size using sound statistical principles.

The appropriate equations for calculating sample sizes are often design dependent. Here, we present a design for simple random sampling. Suppose that d is the allowable error in the sample mean, and the investigator is willing to take only a 5 percent chance that the error will exceed d. In other words, the investigator wants to be reasonably certain that the error will not exceed d. The equation for the sample size is

$$n = \frac{t^2\sigma^2}{d^2} \qquad (3.2)$$

and t is the t statistic for the level of confidence required. For a 95 percent confidence level that the sample mean will not exceed d, $t = 1.96$. Obviously, an estimate of the population standard deviation, σ, is necessary to use this relationship. In many cases, an estimate of σ can be obtained from existing data. When few data are available about σ, it is a good idea to generate a set of tables to develop a sense of the range of samples required.

Suppose, for example, that an investigator wishes to estimate mean pH readings above a wastewater discharge. How many samples are needed to estimate the true mean pH? At the extremes, the investigator guesses that the standard deviation might range between 0.5 and 1.2 pH units. This estimate leads to Tables 3.1 and 3.2:

Table 3.1.—Number of samples needed to estimate the true mean (low extreme).

CONFIDENCE LEVEL	MARGIN OF ERROR (σ=0.5)		
	0.2 pH units	0.5 pH units	1 pH unit
95%	24	4	1
90%	17	3	1

Table 3.2—Number of samples needed to estimate the true mean (high extreme).

CONFIDENCE LEVEL	MARGIN OF ERROR (σ=1.2)		
	0.2 pH units	0.5 pH units	1 pH unit
95%	138	22	6
90%	98	16	4

Note that the number of required samples increases dramatically as the confidence and precision in the estimates increase, and as the population standard deviation increases. As a general rule, the precision of the estimate is inversely proportional to the square root of the sample size. Therefore, increasing the sample size from 10 to 40 will roughly double the precision.

For a fixed precision, changing the required confidence in the estimate from 95 to 99 percent slightly more than doubles the sample size. Equation 3.2 can easily be adopted for binary response variables in which the responses are expressed as proportions or percentages (Cochran, 1963). In addition, for those situations where the number of sampling units is finite, a finite population correction for the sample size is available (Cochran, 1963).

Equations for calculating sample sizes for random, nonrandom, and stratified sample surveys can be found in the literature. They depend on the sample design, the available variance estimates, and whether the environmental assessment has both spatial and temporal components.

Important Rules

Developing a sample design is frequently driven by factors other than statistics and biology. For example, the investigator may be asked to determine a difference between upstream and downstream stations of a municipal treatment plant outfall, long after the suspected impacts began. Even in these cases, creative sampling strategies can help develop the link between the wastewater outfall and downstream impacts. The following rules apply to most environmental assessment scenarios.

- **Rule 1**. Sample designs and their associated analytical techniques can be difficult to conceptualize and implement. Always consult individuals with appropriate training before starting a biocriteria study.
- **Rule 2**. State precisely and clearly the problem under evaluation before attempting to develop a survey design.
- **Rule 3**. Collect samples from a reference site as a basis for inferring impact. In general, the sampling scheme used at the impacted site should be the same as that employed at the reference site.
- **Rule 4**. To the degree possible, use environmental characteristics to minimize the error in the sample estimate. For example, for patchy environments examine the possibility of systematic sampling; for heterogeneous populations, examine the possibility of using stratified random sampling. In all cases, attempt to minimize sample bias by randomly allocating samples (either geographically or temporally across the entire population, or within strata).
- **Rule 5**. For seasonally dependent biocriteria, collect data for several seasons before attempting to determine an impact. For biocriteria that are not seasonally dependent, collect sufficient data to represent the variability in the population.
- **Rule 6**. Collect enough data so that the accuracy and precision requirements associated with using the information are achieved.

BIOLOGICAL CRITERIA
Technical Guidance for Survey Design and Statistical Evaluation of Biosurvey Data

CHAPTER 4 Detecting Mean Differences

Hypothesis testing methods that seek to detect the mean differences arising from two or more independent samples are among the most common statistical procedures performed. However, these procedures are frequently used without regard to some basic assumptions about the data under investigation — which, in some cases, leads to errors in interpretation.

This section describes and illustrates several methods for detecting mean differences. It focuses on (1) cases in which only two means are involved, and (2) situations involving more than two means. It also presents suggestions concerning the use and abuse of means testing procedures.

Cases Involving Two Means

Several scenarios within the biocriteria program require investigators to compare the mean differences between two independent populations. Suppose for example, that we want to use biocriteria in a regulatory setting in the following situation:

A wastewater treatment plant discharges its effluent into a stream at a single point. Upstream of the discharge facility, the stream is in good shape (unaffected by any known sources of pollution). The resource agency has sufficient funds to monitor three stations upstream of the discharge site and a comparable number of streams downstream of the discharge site during the late summer. The agency has chosen to evaluate aquatic life use impairment using benthic species richness.

At each of the six sites, 10 independent measures of species richness were generated by randomly placed ponar grabs over a relatively small spatial area (a sample size of 10 was chosen based on variability estimates generated at a different, but similar site). Sites of comparable habitat quality were chosen for sampling. The upstream sites will serve as a reference condition against which to compare the downstream condition.

In addition to the current survey (i.e., sampling regime, data collection, and interpretation), the regulatory agency has identified an additional upstream site for which it has 10 years of comparable long-term (historical) data. The investigators have no reason to believe that a time component exists in the long-term data. Table 4.1 presents descriptive information associated with the upstream and downstream sites and with the long-term site.

The question for investigators is this: Do the data reveal a downstream effect associated with the wastewater discharge? Several methods are available for assessing the mean differences between the upstream and downstream sites, and each method has both positive and negative aspects.

Random Sampling Model, External Value for σ

Suppose investigators believe that the 30 measures of benthic species richness collected at the upstream and downstream sites can be treated as random samples from appropriate populations. In particular, they

Table 4.1—Descriptive statistics: upstream-downstream measures of benthic species richness.

SITE		*N*	MEAN	STD.	MINIMUM	MAXIMUM	10%–TRIMMED MEAN	MEDIAN ABSOLUTE DEVIATION
Upstream	1	10	10.0	2.3	7.5	14.8	9.7	1.5
	2	10	12.6	2.5	10.3	18.0	12.2	1.3
	3	10	11.2	2.4	7.2	15.1	11.2	1.0
Downstream	4	10	10.4	2.4	6.3	13.7	10.5	1.0
	5	10	7.7	3.7	3.4	14.7	7.4	2.7
	6	10	9.0	1.8	5.6	11.1	9.1	1.5
Historic	7	200	10.4	3.4	0.17	19.4	11.1	2.6
Pooled Data	1–3	30	11.3	2.5	7.2	18.0	11.1	1.0
	4–6	30	9.0	2.9	3.4	14.7	9.0	1.6

believe that the two populations have the same form (i.e., normal distributions with the same variance, σ) but different means, μ_a and μ_b. How can the investigators use statistical theory to make inferences about the effect of the wastewater treatment plant discharge?

If the data were random samples from the populations, with $N_a = 30$ observations from the upstream population and N_b observations from the downstream population, the variances of the calculated averages, Y_a and Y_b would be:

$$V(Y_a) = \frac{\sigma^2}{N_a} \quad , \quad V(Y_b) = \frac{\sigma^2}{N_b} \tag{4.1}$$

Likewise, in the random sampling model, Y_a and Y_b would be distributed independently, so that:

$$V(Y_a - Y_b) = \frac{\sigma^2}{N_a} + \frac{\sigma^2}{N_b} = \sigma^2\left(\frac{1}{N_a} + \frac{1}{N_b}\right) \tag{4.2}$$

Even if the distributions of the original observations had been moderately nonnormal, the distribution of the difference Y_a–Y_b between sample averages would be nearly normal because of the central limit effect. Therefore, on the assumption of random sampling,

$$z = \frac{(Y_b - Y_a) - (\mu_b - \mu_a)}{\sigma\sqrt{\frac{1}{N_a} + \frac{1}{N_b}}} \tag{4.3}$$

would be approximately a unit normal deviate.

Now, σ, the hypothetical population value for the standard deviation, is unknown. However, the historical data yield a standard deviation of 3.4. If this value is used for the common standard deviation of the sampled populations, the standard error of the difference, Y_a-Y_b = 2.3, is

$$\sigma\sqrt{\frac{1}{30} + \frac{1}{30}} = 0.89$$

Based on the robust estimators (trimmed mean difference of 2.1 and median absolute difference of 1.6) the standard error of the difference would be 0.41. If the assumptions are appropriate, the approximate significance level associated with the postulated difference (μ_a–μ_b) in the population means will then be obtained by referring

$$z_0 = \frac{2.3 - (\mu_a - \mu_b)_0}{.89}$$

to a table of significance levels of the normal distribution. In particular, for the null hypothesis (μ_a–μ_b) = 0, z_0 = 2.3/.89 = 2.6, and Pr(z < 2.6) < .005. Again, the upstream/downstream effect seems to be realistic (using the robust estimators, z = 5.1 and Pr[z < 5.1] < .001). Note that we use the z distribution in this example because the population variance is determined from an external set of data that represents the population of interest — an assumption equivalent to assuming that the variance of the population is known (i.e., not estimated).

Random Sampling Model, Internal Value for σ

Suppose now that the only evidence about σ is from the $N_a = 30$ samples taken upstream and the N_b = 30 samples taken downstream. The sample variances are

$$s_a^2 = \frac{\sum(Y_{a1} - Y_a)^2}{N_a - 1} = 6.25$$

$$s_b^2 = \frac{\sum(Y_{b1} - Y_b)^2}{N_b - 1} = 8.41$$

On the assumption that the population variances of the upstream and downstream sites are, to an adequate approximation, equal, these estimates may be combined to provide a pooled estimate of s^2 of this common σ^2. This is accomplished by adding the sums of squares in the numerators and dividing by the sum of the degrees of freedom,

$$s^2 = \frac{\sum(Y_{a1} - Y_a)^2 + \sum(Y_{b1} - Y_b)^2}{N_a + N_b - 2} = 7.52$$

On the assumption of random sampling from normal populations with equal variances, in which the discrepancy [(Y_a–Y_b) – (μ_a–μ_b)] is compared with the estimated standard error of Y_a–Y_b, a t distribution with N_a+N_b–2 degrees of freedom is appropriate. The t statistic in this example is calculated as

$$t = \frac{(Y_a - Y_b) - (\mu_a - \mu_b)}{s\sqrt{\frac{1}{N_a} + \frac{1}{N_b}}} = \frac{2.31}{0.71} = 3.2$$

This statistic is referred to a t table with 58 degrees of freedom. In particular, for the null hypothesis that (μ_a–μ_b) = 0, Pr(t < 3.2) < .001. Again, an upstream/downstream effect seems plausible. Using the robust statistics, a pooled estimate of error can be calculated as the average of the median absolute deviations associated with each data set ([1 + 1.6] / 2 = 1.3). Therefore, the t statistic is 6.3 and the Pr(t < 6.3)< .001. Note that we use the t distribution in this example because the population variance is estimated from the survey data and not assumed to be known.

Testing against a Numeric Criterion

In the preceding sections, hypothesis tests were presented for the two-sample case. Similar tests are avail-

able for testing a sample mean against a fixed numeric criterion (for which an associated uncertainty does not exist). In this case, the *t* statistic can be written as follows:

$$t = \frac{Y - \mu}{s\sqrt{\frac{1}{n}}} \quad (4.4)$$

Here, *s* is the sample standard deviation and μ is the numeric criterion of interest. The probability of a greater value can be found in a *t* table using *n-1* degrees of freedom.

A Distribution-Free Test

In many instances, the assumption that the raw data (or paired differences) are normally distributed does not hold. Even the simplest monitoring design involving the comparison of two means requires either (1) a long sequence of relevant previous records that may not be available or (2) a random sampling assumption that may not be tenable. One solution to this dilemma is the use of distribution free statistics such as the *W* rank sum test (Hollander and Wolfe, 1973). The *W* test is designed to test the hypothesis that two random samples are drawn from identical continuous distributions with the same center. An alternative hypothesis is that one distribution is offset from the other, but otherwise identical. Comparative studies of the *t* and *W* tests indicate that while the *t* test is somewhat robust to the normality assumption, the *W* test is relatively powerful while not requiring normality. In many cases, performing both the *t* and *W* tests can be used as a double check on the hypothesis.

To conduct the *W* test (see Chapter 2), the investigator combines the data points from the samples, but maintains the separate sample identity. This overall data set is ordered from low value to high value, and ranks are assigned according to this ordering. To test the null hypothesis of no difference between the two distributions $f(x)$ and $g(x)$ (i.e., H_0: $f[x] = g[x]$), the ranks of the data points in one of the two samples are summed:

$$W = \sum R_i \quad (4.5)$$

Statistical significance is a function of the degree to which, under the null hypothesis, the ranks occupied by either data set differ from the ranks expected as a result of random variation. For small samples, the *W* statistic calculated in Equation 4.5 can be compared to tabulated values to determine its significance. Alternatively, for moderate to large samples, *W* is approximately normal with mean $E(W)$ and variance $V(W)$:

$$E(W) = \frac{N_a(N_b + N_a + 1)}{2} \quad (4.6)$$

$$V(W) = \frac{N_a N_b (N_b + N_a + 1)}{12} \quad (4.7)$$

$$z = \frac{W - E(W)}{\left[V(W)\right]^{0.5}} \quad (4.8)$$

In the upstream/downstream case that we have been discussing, $E(W) = 1{,}127$, $z = 3.12$, and $\Pr(<z) = 0.0018$.

Evaluating Two-sample Means Testing

Table 4.2 summarizes the advantages and disadvantages of these two-sample means testing procedures. Both of these methods, to one degree of another, involve assumptions of normality, equality of variance, and independence. In all cases, the latter assumption is of greatest concern. Therefore, data with inherent time trends, seasonal cycles, or spatial correlations unrelated to the effect of interest should be carefully scrutinized prior to hypothesis testing using these procedures. Investigators can remove time trends and spatial correlations from the data prior to testing them for mean differences (Reckhow, 1983).

Multiple Sample Case

Hypothesis testing of multiple sample mean differences can be accomplished using both parametric (assumes normality) and nonparametric (no assumption of normality) approaches. The typical parametric approach to multiple means testing falls under the broad class of statistical models and methods called analysis of variance (ANOVA). Nonparametric counterparts include a number of specific tests including, among others, the Kruskal-Wallis rank sum test.

Both the parametric and nonparametric methods can be used with experimental and survey type data. However, the development of these statistical models include many permutations and assumptions and cannot be covered in this text. Instead, a brief discussion of each method is followed by an example of their typical outputs.

Parametric or Analysis of Variance Methods

ANOVA methods are a class of techniques for analyzing experimental data. A continuous response variable, known as the dependent variable, is measured under experimental conditions identified by classification variables known as independent variables or treatments. The variation in response is explained as an effect of the classification variable and random error.

Numerous decisions must be made by the investigator before attempting to use ANOVA procedures.

Table 4.2.—Assumptions, advantages, and disadvantages associated with various two-sample means testing procedures.

REFERENCE DISTRIBUTION	ASSUMPTIONS	ADVANTAGES	DISADVANTAGES	SHOULD CONSIDER FOR USE WHEN:	SHOULD NOT USE WHEN:
External	Past data can provide relevant reference set for observed difference Y_a–Y_b	No assumption of independence of errors. No need for random sampling hypothesis.	Need relevant, lengthy past records. Construction of reference distribution can be tedious	Quality, consistency, and length of data are deemed to represent a healthy ecosystem.	Known impacts to reference site have occurred, or physical and biological differences between the impact and reference site are identified.
Normal distribution with external estimate of σ	Individual observations are as if obtained by random sampling from normal populations with common standard deviation.	Continuous reference distribution that is easy to calculate.	Need to know σ. Need assumption of independence of individual errors coming from random sampling hypothesis.	Quality, consistency, and length of data are deemed to be a sample from a healthy ecosystem. Data transformation may be necessary to achieve normality.	Quality of data is suspect or impacts at the external site are known or suspected.
Normal distribution with internal estimate of σ	Individual observations are as if obtained by random sampling from normal populations with unknown common standard deviation σ estimated by s	No external data needed.	Need assumption of independence of individual errors coming from random sampling hypothesis.	Most commonly used test. Appropriate if normality assumptions hold. If outliers or influential data apprent, consider the use of robust estimators of the mean and variance.	Normality assumptions do not hold. Generally, robust estimators of the mean and variance can reduce the influence of outliers.
Distribution free	Individual observations are as if obtained by random sampling from populations of almost any kind.	Computations are easy. No external data needed. Populations randomly sampled need not be normal.	Need assumption of independence or symmetry of individual errors arising from random sampling hypothesis.	Can be used if normality assumptions are suspect. Can be used to verify results of parametric tests.	No real disadvantage of these tests. In most cases, power of the test is equivalent or near the parametric counterpart.

These decisions include the effects of interest (model specification — one-way designs, two-way designs, and so forth); whether the classification variables are random, fixed, or nested; whether any interactions (nonadditive effects) are present in the data; how to handle unbalanced designs (unequal sample sizes for the various treatments); and the nature of the error term.

As we can see from this list, ANOVA procedures are not simple but require a great deal of thought. In general, the ANOVA model should follow directly from the sample design used to collect the biocriteria data. The following model illustrates a simple one-way, fixed block design like that described in the upstream/downstream case presented here. The overall model for the ANOVA is

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad \textbf{(4.9)}$$

where $Y_{i,j}$ = the j^{th} response for the i^{th} site

μ = the population mean

α_i = the effect of site i on Y

$e_{i,j}$ = the error associated with each observation in the data.

The model assumes that the errors are normally distributed with mean 0 and variance σ^2. Based on the model, any observation is composed of an overall mean (μ), a site effect (α), and a random element (e) from a normally distributed population. Hypothesis testing for the ANOVA model is undertaken by calculating the variance associated with model components (sums-of-square differences around the mean effect). A test statistic is formed by comparing the mean square differences associated with a model component to the mean error term. This statistic is distributed as an F distribution. Table 4.3 presents an example of this variance breakdown for the simple upstream/downstream model.

Table 4.3.— Analysis of variance results for the case study model.

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	Pr>F
Site	5	146.57	29.31	4.51	0
Error	54	350.67	6.49		
Total	59	497.24			

As seen in the table, the effect of site means is an important indicator of the level of benthic species richness. Therefore, it seems a good idea to explore the relationship among the site means as a method of examining a possible gradient of upstream/downstream differences. Several methods are available for testing the differences between site means. In this example, the method of least significant difference (LSD), Duncan's multiple range test, and Tukey's studentized range test are presented. (A review of these and other multiple comparison methods is in the SAS/STAT Guide for Personal Computers.) Tables 4.4 through 4.6 present the results of these multiple comparison tests.

Table 4.4.—Least significant difference multiple comparison test.

GROUPING		MEAN	*N*	SITE
	A	12.6	10	2
B	A	11.2	10	3
B	A	10.4	10	4
B		9.9	10	1
B	C	8.9	10	6
	C	7.6	10	5

Table 4.5.—Duncan's multiple comparison test.

GROUPING		MEAN	*N*	SITE
	A	12.6	10	2
B	A	11.2	10	3
B	A	10.4	10	4
B	C	9.9	10	1
B	C	8.9	10	6
	C	7.6	10	5

Table 4.6.— Tukey's multiple comparison test.

GROUPING			MEAN	*N*	SITE
	A		12.6	10	2
B	A		11.2	10	3
B	A	C	10.4	10	4
B		C	9.9	10	1
B		C	8.9	10	6
		C	7.6	10	5

In the above tables, sites within a specified grouping are not different at the $\alpha = 0.05$ level of significance.

Nonparametric or Distribution Free Procedures

Distribution free methods for testing multiple sample means are available in much the same format as for parametric tests. The Kruskal-Wallis rank sum test (one-way design) and the Friedman rank sum test (two-way design) are frequently used when the normality assumptions do not hold (see Hollander and Wolfe [1973] for a review of these methods). Multiple comparison methods based on the individual rank scores for each site are available.

Again, the investigator must develop the model to match the experimental design. In the upstream/downstream comparisons of benthic species richness, the Kruskal-Wallis test with a simple one-way model results in a chi-square statistic of 16.38 (Pr < chi-square = 0.006). Again, the upstream/downstream sites appear to differ in the measured biocriteria. Results of the multiple comparison tests using ranks were similar to those presented in the ANOVA model.

A Test for Broad Alternatives

Frequently, investigators are faced with situations in which tests for mean differences or variance differences are not sufficient. For example, investigators

may realize that smaller fish are more sensitive to a pollutant than larger fish. In such cases, simple testing for mean differences (in which the mean is calculated without regard to size class) between reference and impacted sites may not suffice. Instead, the measure of toxic effect will be better reflected through changes in the distribution of fish caught at the two sites. Examining the differences in distribution functions among sites may be a more sensitive way to detect effects than relying on population estimates such as the mean and variance.

Statistics designed to detect broad classes of alternatives, as in the scenario presented here, are distribution free tests (i.e., they do not rely on normality assumptions), although they do have parametric counterparts. For a single sample, goodness-of-fit tests to gage the correspondence between an empirical distribution function of observations and a specific probability model or distribution (e.g., normal or lognormal) may be useful. These tests can also be conducted using the chi-square statistic (see Snedecor and Cochran, 1967).

The Kolmogorov-Smirnov Two-Sample Test

Within the biocriteria program investigators will frequently be challenged to evaluate a broad range of differences between two or more populations. The Kolmogorov-Smirnov (KS) two-sample test is easy to implement and can be used to evaluate the relationship between two distribution functions. This test provides graphic and statistical evaluations of two sets of data.

The KS two-sample test involves the development of two cumulative distribution functions (CDFs) to test the hypothesis that each sample was taken from the same population. The test is based on the difference between the empirical distribution functions. The largest difference between the two functions, D_{max}, forms the basis for the test statistic. D_{max} is the maximum vertical distance at any horizontal point between the two CDFs (Fig. 4.1).

To generate a CDF for an individual sample, the data are ordered from lowest to highest, and the rank order of each point determined. Dividing each rank by the sample size results in a cumulative distribution function ranging from 0 to 1 (or 0 to 100 percent, if multiplying by 100). The two samples need not have the same number of observations. Tabled values of the test statistic are available for various sample sizes (Hollander and Wolfe, 1973). The test is both one-sided and two-sided. For the benthic species rich-

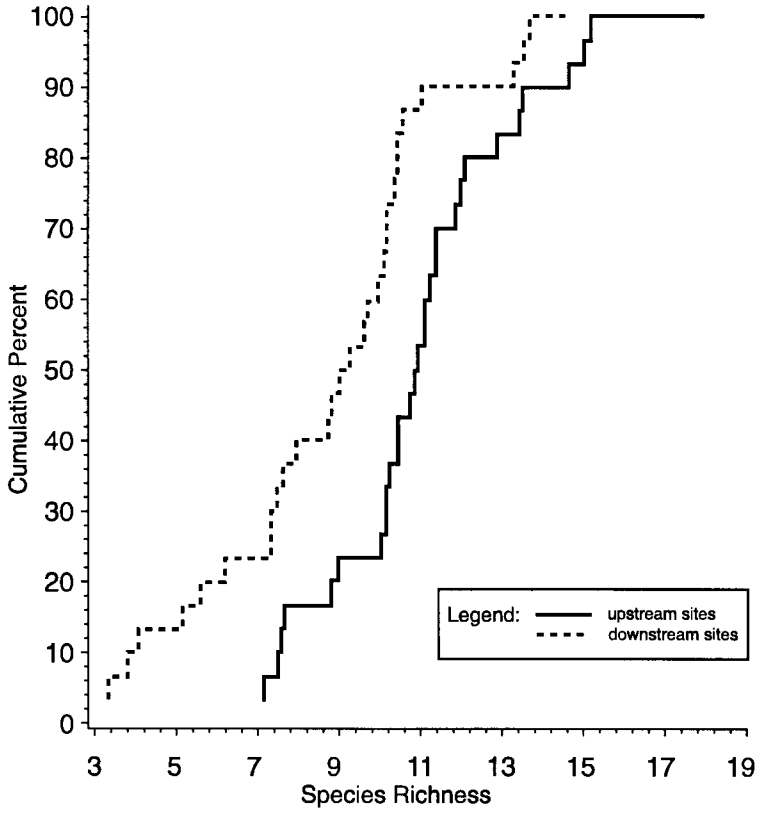


Figure 4.1—Cumulative distribution functions of upstream and downstream sites.

ness example shown here in Figure 4.1, D_{max} is 0.433 (43.3 percent) which occurred at a species richness value of 10.6. The null hypothesis is rejected with a Type I error rate of 0.0072.

Relationship of Survey Design to Analysis Technique

Table 4.7 outlines the relationship between means testing techniques and selected survey designs as described in earlier sections. As a general rule, the data analysis techniques are driven by the survey design. The principle decision points are the number of sites, the available sample size, and the presence or absence of reference sites. However, investigators should not be constrained by the survey design. Data exploration, using any technique that fits the data, is encouraged and can provide insightful results.

Table 4.7.—Survey design and analysis techniques.

SURVEY DESIGN	MEAN DETECTION METHOD
Upstream/downstream: random sampling at single sites using current survey data	*t*-test using an internal value of the variance; Wilcoxon test; with large data sets, a KS two-sample test may be appropriate
Upstream/downstream: random samplings at multiple sites using current survey data	One-way ANOVA using an internal value of the variance; KS two-sample test on merged upstream and downstream data; Kruskal-Wallis rank sum test
Upstream/downstream: random sampling within spatial or temporal strata with one or more sites	Two-way (or more complicated) ANOVA tests; Friedman rank sum test (and other more complicated nonparametric tests)
Impact site data with large off-site external data; for example when determination of impact is not clearly definable or no good upstream reference condition available	External reference distribution tests including the two-sample KS test; *t*-test with external estimate of the variance
Systematic sampling such as random sampling along a transect or nodes of a grid	ANOVA, *t*-tests with internal estimates of the variance, and possibly distribution tests (also note that such designs may be subjected to techniques that demonstrate geographical trends and patterns such as kriging and GIS methods)
Regionally impacted sites with one or more reference sites	Two-sample KS test

BIOLOGICAL CRITERIA
Technical Guidance for Survey Design and Statistical Evaluation of Biosurvey Data

CHAPTER 5 Discussion and Examples

In the previous four chapters, standard statistical methods were presented, discussed, and illustrated with simple examples. Those methods and examples represent conventional analyses or situations in which sample sizes are relatively large so that hypothesis testing is essentially straightforward. The analyses were motivated by available, commonly applied methods, and the examples were structured to fit the methods. The purpose was to provide background statistical guidance, with examples.

In this chapter a different approach is taken. Here, typical problems involving biosurvey data are the starting points, and statistical methods for analysis and hypothesis testing are proposed and applied specifically to the problem. In some cases, hypothesis testing is possible; in others, the small sample size may limit statistical inference. In the latter situation, the investigator may consider design changes so that different statistical analyses can be undertaken with biosurvey data in the future.

We begin with a general discussion of the importance of small sample size and briefly examine judgmental and statistical options for small sample size, followed by examples of hypothesis testing with small samples. The chapter concludes with "rules of thumb."

Working with Small Sample Sizes

The conventional methods for statistical hypothesis testing and interval estimation presented in chapters 1 through 4 work best under conditions that do not always exist with biosurvey data. The common approaches based on an underlying normal probability model are clearly not essential; distribution-free methods are versatile and effective. Still, virtually all confirm-atory analyses (i.e., those concerned with hypothesis testing and interval estimation) require estimation of a "location" statistic that is the quantity of interest (e.g., a mean, median, or quartile), and they also require estimation of a variability statistic (e.g., a standard error) that indicates the spread of values for the location statistic.

An example of a desirable scenario for confirmatory statistical analysis was described in Chapter 2. Data must be available from the sites of direct interest in the assessment, and sample sizes must be large enough for hypothesis testing. If the site-specific data are inadequate (less than two, which would prevent direct calculation of a sample variance), or too small, (e.g., less than five, which would make the calculated sample variance quite uncertain), then alternatives to statistical testing or intervals are possible, but these alternatives are apt to include additional conditions or assumptions beyond those required in conventional analyses.

For example, a single sampling might yield a point estimate for IBI downstream of a wastewater discharge, but provide no measure of variability. If historic data exist on IBI at other impacted sites, then it is reasonable to assume that the variability in the historic data can be used as the variability measure for testing at the site of interest. If, on the other hand, the historic data analysis includes an IBI regression based on predictors, such as watershed area and physical habitat quality, then the standard error for this regression is the appropriate variability measure. The key feature of these hypothetical examples is that other, relevant information exists that the investigator believes can be used to estimate statistics for the site of interest.

In the absence of historic data for statistical estimation (usually for the estimate of variability), hypothesis testing and interval estimation may still be possible if the scientist is prepared to make certain assumptions. For example, suppose that an aquatic biologist is confident that he or she can estimate the variability in IBI in impacted streams based on experience and knowledge of the literature. This estimate could provide the necessary variability measure, but it is obviously conditional on the judgment of the biologist.

None of the approaches presented in this document are without assumptions; even the example in Chapter 2 includes the assumption that the sample data adequately reflect the true situation. Judgment-based estimates of statistics require a different assumption, namely, the assumption that the investigator's judgment is good.

The most serious difficulty in the application of interval estimation and hypothesis testing for biosurvey data is the small sample size associated with many biological surveys. The strength of inferences from statistical analysis is tied to sample size. If expert judgment is not available or not acceptable, then sample size must be large; otherwise, statistical testing is either not possible or not particularly useful. But how large is "large enough"? There is no single, correct answer to that question. As a rule, the stan-

dard error drops according to the square root of the sample size; thus, the answer to the question depends on the error level that is acceptable in the problem under study.

In general, sample sizes greater than 10 are usually desirable, and sample sizes smaller than five may prevent meaningful statistical testing. In addition, since standard error may be expected to drop with the *square root* of the sample size, there are diminishing returns as sample size grows larger.

What can be done when sample size is too small and expert judgment is either not available or not acceptable? Any amount of data or evidence can indicate an effect (or the absence of an effect), and this information can be described in text, presented in tables, or displayed in graphs. However, in the case of very small samples, it is important to emphasize that the analysis is descriptive and not confirmatory. Alternatively, if the investigators have data on biological and chemical indicators of impairment and criteria for each of the indicators, then it may still be possible to test effects across indicators.

Suppose there is no sample size estimate — only an estimate of variability based on expert judgment. How can statistical testing be completed? We actually have some well-established approaches to elicit judgment-based quantities and error estimates, along with an effective number of degrees of freedom (Meyer and Booker, 1991). Alternatively, the scientist may simply summarize test results in a table with sample size (or degrees of freedom) and test results (e.g., *p*-values) given for a range from small to large samples. In some cases, the conclusion may not depend on the effective sample size; in others, sample size may be critical, which places more importance on the goodness of the judgmental assessment.

Assessments Involving Several Indicators

Suppose that sampling has occurred at a stream site at which environmental degradation is suspected, but the sample size for any single indicator is too small for hypothesis testing. For each indicator, the state has established an impairment criterion; thus, the results of sampling could be presented either as a measurement (e.g., dissolved oxygen concentration) or as success or failure in meeting the state's criterion. Each of the indicators is expected to provide an independent measure or assessment of environmental degradation; therefore, several indices cannot be separately included in the analysis if they are based on the same underlying measurements.

As an example, Table 5.1 presents three biological indices, the IBI, ICI, and Iwb based on sampling at a single site on three different dates. The state biocriteria are also given. It is assumed that the two-month period between samplings results in temporal independence between the samples.

Table 5.1—Biological Indices and biocriteria

DATE	IBI	ICI	Iwb
June 15	43(1)	38(1)	9.4(1)
August 15	39(0)	38(1)	8.7(1)
October 15	42(1)	36(1)	8.3(1)
Biocriteria	40	35	8.5

Since we have only a single estimate per date on each index, and only three data points per date and per index, statistical inference opportunities are limited. We can, however, treat the nine index estimates in Table 5.1 as nine independent measures by which to assess the underlying condition of biologic impairment, based on biocriteria violations. The indices in Table 5.1 are recorded as a 0-1 variable, in parentheses, indicating attainment (1) or violation (0) of each biocriterion. Next, these nine 0-1 data points can be subjected to statistical analysis to determine the overall biologic impairment reflected in the aggregate of the three indices.

First, calculate the proportion of violations (p) in the sample as an estimate for the probability of biologic impairment at the site:

$$\hat{p} = \frac{2}{9} = 0.222$$

$\hat{p}$ is a point estimate that is uncertain due to natural variability and measurement error. We can calculate a confidence interval for $\hat{p}$ or test the hypothesis that $\hat{p}$ is less than a specified critical value. Once it is calculated, a confidence interval or a percentile could serve as a cutoff point indicative of biological impairment. For example, one might define impairment as more than 50 percent violations. As a variation on that idea, Rankin and Yoder (1990) selected the 75th percentile in a histogram of sample IBI deviations (from the mean value) to be the limit of tolerable variation.

Confidence intervals for $\hat{p}$ can be determined using binomial tables or graphs like those presented in Hahn and Meeker (1991), or using Table 1.4.1 in Snedecor and Cochran (1967). For example, the two-sided 90 percent confidence interval for this example (based on Table A.23a in Hahn and Meeker) is

$$0.041 \leq p \leq 0.550$$

If instead of binomial tables, the large sample normal approximation is used (see Snedecor and

Cochran, 1967), the two-sided 90 percent confidence interval is

$$\hat{p} - 1.645\sqrt{(\hat{p})(1-\hat{p})/n} \le p \le \hat{p} + 1.645\sqrt{(\hat{p})(1-\hat{p})/n}$$

which, for this example is

$$\frac{2}{9} - 1.645\sqrt{\frac{2}{9}(\frac{7}{9})/9} \le p \le \frac{2}{9} + 1.645\sqrt{\frac{2}{9}(\frac{7}{9})/9}$$

$$0 \le p \le 0.450$$

Clearly, the large sample normal approximation is not appropriate for this small sample.

The binomial confidence interval for $\hat{p}$ is quite large as a consequence of the small sample size; this illustrates how small samples can hamper rigorous statistical inference. Nevertheless, the information in even a small number of samples can be expressed graphically (e.g., using a histogram) or in statistics characterizing center and dispersion. Following Rankin and Yoder, a percentile can be selected from the histogram to serve as a biocriterion.

Note that this percentile reflects variability in the sample, but not strength of evidence as conveyed in sample size or degrees of freedom. The advantage in using a confidence interval rather than an empirical distribution percentile is that the sample size is incorporated in the confidence interval. Thus, more information, expressed as a larger sample size, translates properly to a smaller confidence interval (and indicates greater strength of evidence).

In many applications, intervals may be one-sided, since only one side or bound is of interest. In this example, the two-sided 90 percent confidence interval upper cutoff of 0.55 is the one-sided upper bound on the 95 percent confidence interval. From this information, an infinite number of impairment criteria are possible. One option is to require that $\hat{p} = 0.5$ be outside the upper 95 percent confidence interval for attainment; this could be interpreted as indicating only a slight possibility, a 50/50 chance, of overall biocriteria violation. With that impairment criterion, analysis of the data in Table 5.1 leads to failure to achieve attainment. This conclusion would be reversed, even if the 2/9 biocriteria index violation rate continued, if more samples were collected leading to a tighter confidence interval. The conclusion would also be different for roughly the same sample size if the frequency of biocriteria index violation were lower.

Regional Reference Data

Bioassessment data on regional conditions (e.g., regional reference sites) may sometimes be used with small sample sizes, or even with a single sample, to go beyond a point estimate of status. Consider, for example, the information presented in Yoder (1991). He compared the assessments from the application of the Ohio narrative macroinvertebrate criteria from 1979 through 1986 with a calculated ICI score. In this study, about 400 sites were rated using both narrative macroinvertebrate criteria and calculated ICI; and the two ratings were then compared for each of the sites.

Yoder expressed this comparison using three ICI distributions: the ICI scores for the sites labeled "good/exceptional" based on the narrative criteria, the ICI scores for the "fair" sites, and the ICI scores for the "poor/very poor" sites (see Yoder, 1991, Fig. 7). Yoder argued that the ICI scores are more reliable than are the classifications based on the narrative criteria, and he employed point cutoffs between classes (ICI = 35 between good and fair; ICI = 13 between fair and poor).

If, unlike Yoder, we take the ICI distributions for each of the three classes as reference distributions, then we can use the classification rules typically employed with discriminant analysis (see Flury and Riedwyl, 1988) to estimate the probability that any new sites belong in each class. To do this, we must make a distributional assumption concerning the probability model that describes the ICI within each class. As a rule, it is assumed that this distribution (of ICI) is normal (within each of the three classes), with mean and variance estimated based on the sample (ICI) values.

As another example of small sample size data sets, imagine that repeated IBI measurements are taken both from a reference site, and from a site with known anthropogenic pollutants. Data from each of the sites are analyzed, and their respective distribution functions are created. Such a case is presented in Figure 5.1. Here, the IBI sampling distributions for each site are roughly shaped as a normal distribution.

Assume, further, that a single IBI measurement from a third site is generated. This single measurement is shown on Figure 5.1 as a solid vertical line. Does the investigator have enough information to categorize the site as impacted or not impacted? Visual examination of the figure shows that the third site IBI measurement lies in an area of substantial overlap between the impacted and reference site distributions. Therefore, given that the sampling error of the third site is unknown, it is difficult to assess with confidence whether the third IBI measurement is consistent with either the reference or impacted sites.

At this point, the investigator would be best served if he or she gathered additional IBI measurements. If, on the other hand, the single-sample IBI measurement had been in the tail of either distribution (say, an IBI of 20 or 55), then the investigator could have classified the third site appropriately. In

making this classification, the investigator would have noticed that little overlap of the distributions occurs in the extreme tails of the impacted and reference site distributions.

Using Background Variability Measures

In the previous section, the Ohio ICI biocriteria were identified as point values between classes (e.g., ICI = 35 is the warmwater habitat criterion separating "good/exceptional" from "fair"). When a single ICI determination is available from a new site, the Ohio criteria can be used to classify the site, ignoring uncertainty. Beyond that, if it is assumed that the Ohio ICI classification scheme is fixed and certain, and if a reliable estimate of site ICI variability is available, then the classification based on a single ICI value can be assessed using a hypothesis test.

In situations with only a single estimate of a bioindicator, collateral information must be obtained to provide the estimate of variability. There are several potential measures of site bioindicator variability that might be suitable; Rankin and Yoder's (1990) discussion presents several informative graphs to show, for example, that the IBI coefficient of variation drops as IBI increases (Rankin and Yoder, 1990, Fig. 2), and IBI coefficient of variation increases slightly as drainage area increases (ibid., Fig. 7).

Knowledge and judgment can be quite helpful in selecting the variability estimate. For example, if it is believed that the site bioindicator variability is roughly constant within a specified category, then a calculated estimate of variability for the bioindicator within the appropriate class can be used as the variability measure for the site of interest. Categories may be selected on any criterion (e.g., ecoregion, IBI range) that is scientifically plausible and leads to an acceptably large overall sample size for variability estimation.

Rankin and Yoder's graphs suggest that, while the IBI coefficient of variation changes with selected categories (IBI range), the IBI standard deviation may be roughly constant across IBI classes and across ecoregions. A median standard deviation between 4 and 5 appears to be quite consistent in the graphs. Based on this collateral information, it is assumed that site-specific IBI in Ohio, under constant conditions (i.e., no change in site factors that determine IBI), has a standard deviation of 4.5.

Here is an example of how this estimate is used. Assume that the single IBI measurement shown in Figure 5.1 (IBI = 35) was taken in Ohio under the conditions described. Since the sampling program in Ohio is quite large, 4.5 is effectively the true standard deviation for all sites; thus, with a single sample, it may be concluded that the standard error for the mean value (IBI = 35) is also 4.5. To determine whether the sample is taken from the reference or impacted distribution, assume that 18 IBI samples were taken at the reference and impacted sites, and that the following statistics are calculated:

Reference site sample mean = 42, sample standard deviation = 5;

Impacted site sample mean = 27, sample standard deviation = 8.

Then, a two-tailed t test using Equation 2.1b (see Chapter 2) evaluating the null hypothesis that the means are the same will result in the following:

t = 1.43, for the hypothesis that the reference site mean is equal to the mean of the third site mean; and

t = 1.245, for the hypothesis that the impacted site mean is equal to the third site mean.

Based on this information, the investigator has some evidence that the sample collected from the third site is closer to the impacted site mean than to the reference site mean. However, as conveyed by the similar t statistic results, the confidence in this conclusion is relatively weak.

Final Suggestions for Small Sample Sizes

The discussion and examples in this chapter, while intended as useful, general guidance, are not firmly rooted in statistical theory and hence not always to be followed. Rather, they reflect our experience and observations. Further, they concern the real situations that biologists confront — situations that do not conform to well-established statistical procedures. However difficult and awkward for statistical analysis, the problems must be addressed. With this caveat, the following concluding comments summarize the discussion and examples presented here:

1. If the sample size is 1, a measure of variability may still be obtained using expert judgment or other data. If no variability measure can be justified, then descriptive statistics may be the extent of the analysis (i.e., no interval estimation or hypothesis testing).

2. If the sample size is more than 1 but still small (perhaps 5 or fewer), then it is possible to use the sample to estimate variability for interval estimation or hypothesis testing. However, the intervals may be very large and the tests not

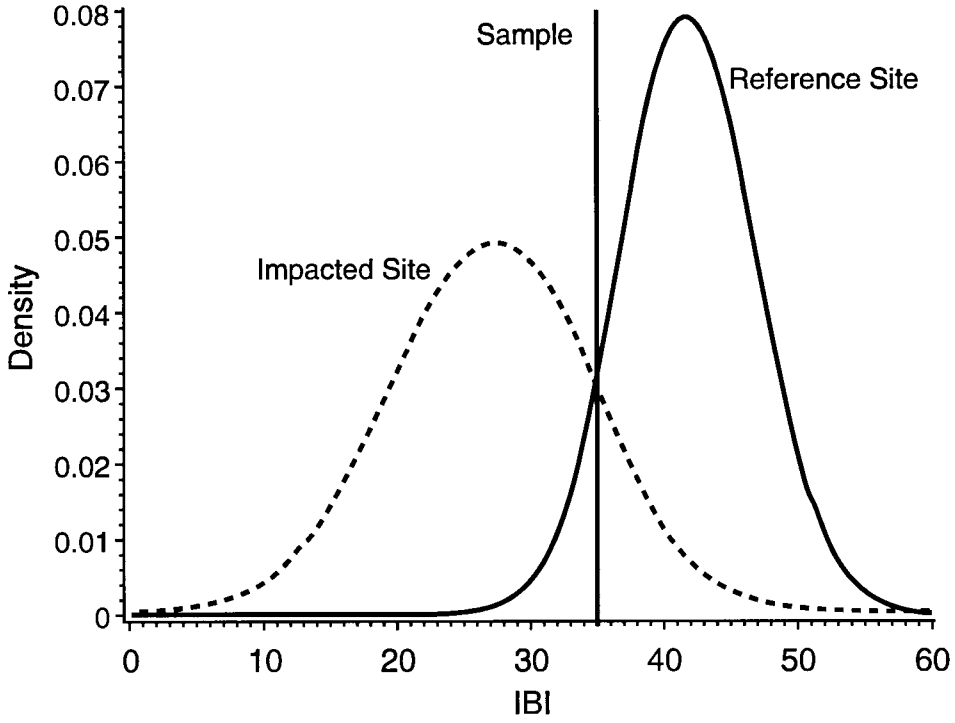


Figure 5.1—IBI Distributions for reference and impacted sites

very powerful, because small sample size means that the strength of evidence is weak.

3. Situations may exist with more than a single estimate of variability. Perhaps one estimate will be based on data and a second estimate on expert judgment. In that case, the two estimates of variance can be pooled, using an estimator like that in Chapter 4's "Reference Distribution Based on Random Sampling Model, Internal Value for σ." A difficulty in pooling when a judgmental estimate of variance is involved is determination of the degrees of freedom for the judgmental variance estimate. Perhaps the best approach is to make a reasoned guess as to how much information the judgment contains with respect to samples (the "effective sample size"):

   (a) if the judgment is highly uncertain, assign it a small number of degrees of freedom (perhaps 2-5),

   (b) if there is more confidence in the judgment, assign the judgment estimate 5+ degrees of freedom.

   If the conclusions from this analysis are not particularly sensitive to the exact choice of the effective sample size for the judgmental estimate, then inferences may be made with some confidence. If, however, the conclusions are sensitive to this choice, then the best approach may be to obtain additional information before drawing final conclusions.

Decision Analysis and Uncertainty

In the preliminary approach presented here we have advocated the use of classical statistical hypothesis testing to summarize data concerning biological criteria. We assume that a decision and succinct conclusions based on the data are needed. However, alternatives to hypothesis testing may be appropriate in certain situations. For example, statistical and graphic summaries (e.g., confidence intervals, bivariate plots) may be used to summarize and present information when the investigator believes that a classical hypothesis test based on a single parameter is too brief or that more evidence should be presented.

An alternative is to recast the hypothesis testing problem using a decision analytic framework. Decision analysis (Raiffa, 1968; Reckhow, 1984) begins with the scientific base summarized in the hypothesis test and incorporates the consequences (e.g., costs and benefits) of possible decisions. In an informal analysis, a decision analytic approach may be undertaken by the decision maker if a desired outcome of management action is "to hedge away" from large adverse consequences or losses. Informal considerations and hedging may be most effectively undertaken in an a priori assessment of costs and

benefits, which then becomes a primary basis for choosing between various levels of test significance. Thus, if it seems likely that biological degradation can be avoided, then the decision maker may request that the biologist set the significance level for testing (e.g., that H_0 has no impact) relatively high (e.g., at 0.10 or 0.20). Alternatively, if cleanup costs are high relative to benefits, then the test significance level (for H_0 has no impact) could be set relatively low (e.g., at 0.01 or 0.005).

Suppose that a measure of biological integrity is tested for upstream-downstream differences surrounding wastewater treatment plant discharges from small treatment plants (less than 5 million gallons per day) throughout the state. If the per person cost to upgrade the treatment level for small communities is generally quite high, and the benefits to be derived from biological improvements are generally low (relative to the organisms affected and typical uses of the streams), hedging away from high cost may be informally undertaken by setting the significance (or "action") level of the test quite low (e.g., 0.01 or 0.005). Additional study of biological degradation, costs, and benefits would be triggered only if an upstream-downstream test result was significant at this level.

Hedging away from large losses is an option precisely because of scientific uncertainty. If there were no scientific uncertainty about biological degradation, then the analysis would always focus on costs and benefits, and the management option with the highest net benefits would be selected. On the other hand, if scientific uncertainty is extreme, an appropriate strategy may be either to hedge farther from large adverse consequences or to seek more information, if possible, to reduce scientific uncertainty before new management action is adopted.

In more formal applications, decision analysis may be used to combine uncertain scientific information on biocriteria (expressed probabilistically) with an overall measure of net benefits or use associated with management actions. This approach is most effective in a Bayesian context; Reckhow (1984) presents a simple example applied to lake eutrophication management. However, comprehensive Bayesian decision analysis is apt to be prohibitively expensive (in terms of human resources and cost) for all but the most critical and consequential problems.

One outcome of data analysis may be that the decision maker will desire more information before implementing new management actions. In formal decision analysis, a value of information calculation should be made to help one determine the wisdom of immediate action versus additional data collection and analysis. In informal analysis, one should consider how useful new information would be if action has to be deferred pending its arrival.

The outcome of hypothesis testing is a statistical summary of evidence on biological degradation. It does not establish cause and effect, although a well-designed test may associate degradation with a candidate cause. The strength of causal conclusions depends on a number of factors including a priori scientific knowledge and field observation. Scientific support for management actions is greatest when the observation of degradation is accompanied by documentation of a causal relationship.

In most cases, environmental management decisions reflect a certain limited understanding of causal connections and a certain degree of observational evidence that is more statistical in nature. This combination is a reasonable basis for decision; in fact, it would be unreasonable to expect detailed causal knowledge in support of every decision. However, as management actions are undertaken and biological response is observed after the fact, more observational evidence may be gathered to support earlier decisions.

REFERENCES

Andrews, D.F., P.J. Bickel, F.R. Hampel, P.J. Huber, W.H. Rogers, and J.W. Tukey. 1972. Robust Estimates of Location. Princeton University Press, Princeton, NJ.

Barnett, V., and T. Lewis. 1984. Outliers in Statistical Data. 2nd edition. John Wiley and Sons, Chichester, UK.

Blalock, H.M. Jr. 1972. Social Statistics. McGraw-Hill, New York.

Box, G.E.P., J.S. Hunter, and W.G. Hunter. 1978. Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building. John Wiley and Sons, New York.

Cochran, W.G. 1963. Sampling Techniques. John Wiley and Sons, New York.

Cochran, W.G., and G. M. Cox. 1957. Experimental Designs. John Wiley and Sons, New York.

Conover, W.J. 1980. Practical Nonparametric Statistics. 2nd edition. John Wiley and Sons, New York.

Dixon, W.J., and J.W. Tukey. 1968. Approximate behavior of the distribution of Winsorized t (trimming/Winsorization 2). Technometrics 10:83-98.

Flury, B., and H. Riedwyl. 1988. Multivariate Statistics, a Practical Approach. Chapman and Hall, London.

Gilbert, R.O. 1987. Statistical Methods for Environmental Pollution Monitoring. Van Nostrand Reinhold, New York.

Green, R.H. 1979. Sampling Design and Statistical Methods for Environmental Biologists. John Wiley and Sons, New York.

Hahn, G.J., and W.Q. Meeker. 1991. Statistical Intervals. John Wiley and Sons, New York.

Hampel, F.R., E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel. 1986. Robust Statistics: The Approach Based on Influence Functions. John Wiley and Sons, New York.

Hill, M.A., and W.J. Dixon. 1982. Robustness in real life: a study of clinical laboratory data. Biometrics 38:377-96.

Hollander, M., and Wolfe, D.A. 1973. Nonparametric Statistical Methods. John Wiley and Sons, New York.

Hunsaker, C.T., and D.E. Carpenter, eds. 1990. Environmental Monitoring and Assessment Program: Ecological Indicators. EPA/600/3-90/060. Off. Res. Dev., U.S. Environ. Prot. Agency, Washington, DC.

Horn, P.S., P.W. Britton, and D.F. Lewis. 1988. On the prediction of a single future observation from a possibly noisy sample. The Statistician 37:165-72.

Huber, P.J. 1981. Robust Statistics. John Wiley and Sons, New York.

Hurlbert, S.H. 1984. Pseudoreplication and the design of ecological field experiments. Ecolog. Monogr. 54:187-211.

Iglewicz, B. 1983. Robust scale estimators and confidence intervals for location. Pages 404-31 *in* D.C. Hoaglin, F. Mosteller, and J.W. Tukey, eds., Understanding Robust and Exploratory Data Analysis. John Wiley and Sons, New York.

Kmenta, J. 1986. Elements of Econometrics. 2d ed. Macmillan, New York.

Linthurst, R.A., et al. 1986. Population Descriptions and Physico-Chemical Relationships. Vol 1 of Characteristics of Lakes in the Eastern United States. EPA/600/4-86/007a. U.S. Environ. Prot. Agency, Washington, DC.

Meyer, M. A., and J.M. Booker. 1990. Eliciting and Analyzing Expert Judgement: A Practical Guide. Academic Press, London.

Miller, R.G. Jr. 1986. Beyond ANOVA: Basics of Applied Statistics. John Wiley and Sons, New York.

Morgan, M.G., and M. Henrion. 1990. Uncertainty. Cambridge University Press, UK.

Mosteller, F., and J.W. Tukey. 1977. Data Analysis and Regression: A Second Course in Statistics. Addison-Wesley, Reading, MA.

Ohio Environmental Protection Agency. 1988. The Role of Biological Data in Water Quality Assessment. Vol. 1 of Biological Criteria for the Protection of Aquatic Life. Div. Water Qual. Monitor. Assess., Columbus, OH.

Raiffa, H. 1968. Decision Analysis. Addison-Wesley, Reading, MA.

Rankin, E.T., and C.O. Yoder. 1990. The nature of sampling variability in the index on biotic integrity (IBI) in Ohio streams. EPA-905-9-90/005. Pages 9-18 *in* Proc. 1990 Midw. Pollut. Meet., Chicago, IL.

Reckhow, K.H. 1979. Techniques for exploring and presenting data applied to lake phosphorus concentration. Can. J. Fish. Aquat. Sci. 37(2):290-94.

———. 1984. Decision theory applied to lake management. Pages 196-200 *in* Proc. Fourth Ann. Conf. N. Am. Lake Manage. Soc., City and State?

Reckhow, K.H., and S.C. Chapra. 1983. Data Analysis and Empirical Modeling. Vol 1 of Engineering Approaches for Lake Management. Butterworth Pubs., Boston, MA.

Reckhow, K.H., and C. Stow. 1990. Monitoring design and data analysis for trend detection. Lake Reserv. Manage. 6(1):49-60.

Reckhow, K.H., K. Kepford, and W. Warren-Hicks. 1993. Statistical Methods for the Analysis of Lake Water Quality Trends. EPA 841-R-93-003. U.S. Environ. Prot. Agency, Washington, DC.

Rey, W.J.J. 1983. Introduction to Robust and Quasi-Robust Statistical Methods. Springer-Verlag, Berlin.

Rocke, D.M. 1983. Robust statistical analysis of interlaboratory studies. Biometrika 70:421-31.

Rocke, D.M., G.W. Downs, and A.J. Rocke. 1982. Are robust estimators really necessary? Technometrics 24(2):95-101.

Snedecor, G.W., and W.G. Cochran. 1967. Statistical Methods. 6th ed. Iowa State University Press, Ames.

Staudte, R.G., and S.J. Sheather. 1990. Robust Estimation and Testing. John Wiley and Sons, New York.

Stevens, D. 1989. Field sampling design. *In* W. Warren-Hicks and B. Parkhurst, eds., Ecological Assessment of Hazardous Waste Sites: A Field and Laboratory Reference. EPA/600/3-89/013. Environ. Research Lab., U.S. Environ. Prot. Agency, Corvallis, OR.

Stigler, S.M. 1977. Do robust estimators work with real data? Ann. Stat. 5(6):1055-98.

Tukey, J.W. 1960. A survey of sampling from contaminated distributions. Pages 448-85 *in* I. Olkin, ed., Contributions to Probability and Statistics, Stanford University Press, Stanford, CA.

———. 1977. Exploratory Data Analysis. Addison Wesley, Reading, MA.

Tukey, J.W., and D.M. McLaughlin. 1963. Less vulnerable confidence and significance procedures for location based on a single sample: trimming/Winsorization. Sankhya A. 25:331-52.

U.S. Environmental Protection Agency. 1990. Biological Criteria, National Program Guidance for Surface Waters. EPA-440/5-90-004. Off. Water Reg. Stand., Washington, DC.

U.S. Government Printing Office. 1988. The Clean Water Act as amended by the Water Quality Act of 1987. Pub. L. 100-4, Washington, DC.

Warren-Hicks, W.J., and J. Messer. 1990. Using Biological Indices to Measure Ecological Condition in Regional Resources. Draft Rep. Prep. for Atmos. Res. Exposure Assess. Lab., Research Triangle Park, NC.

Williams, B. 1978. A Sampler on Sampling. John Wiley and Sons, New York.

Wonnacott, T.H., and R.J. Wonnacott. 1977. Introductory Statistics. John Wiley and Sons, New York.

Yoder, C.O. 1991. Answering some concerns about biological criteria based on experiences in Ohio. Pages 95-104 *in* G.H. Flock, ed., Water Quality Standards for the 21st Century. Proc. Off. Water, U.S. Environ. Prot. Agency, Washington, DC.

Yuen, K.K., and W.J. Dixon. 1973. The approximate behaviour and performance of the two-sample trimmed t. Biometrika 60:369-74.

@REF =